Iterated learning of language distributions

B001285

MSc Evolution of Language and Cognition

The University of Edinburgh

2011

Contents

List of Figures

Abstract

This dissertation presents the results of a series of simulations intended to expand the findings of Burkett and Griffiths (2009, 2010), whose model is shown to make a number of assumptions that may be unrealistic with regard to human language learners. These assumptions are modified to create a number of more realistic scenarios. A series of simulations shows that the concentration parameter α continues to affect the outcome of iterated learning with Bayesian learners in these new scenarios. To overcome the need for the concentration parameter to be specified by the modeller, a model is presented where agents learn a complex hypothesis composed of both a distribution of languages within a population and the appropriate value for α. The outcome of the simulations based on this model are inconclusive but do hint at the possibility of α being affected by iterated learning, potentially enabling learners to acquire a complex hypothesis.

Acknowledgements

I would like first of all to thank my advisor, Dr. Kenny Smith, for his invaluable support and advice as well as his encouragement and seemingly infinite patience when things were not going as planned. I would also like to thank my children, Ilja and Malin, for reminding me on a daily basis that although our ability to learn two or three languages is amazing, it is also the most natural thing in the world. Finally, my wife, Magi, had far more faith in me than I did; I cannot begin to thank her enough for her love and support.

CHAPTER 1

Introduction

This dissertation presents the results of a series of simulations that expand on the findings of Burkett and Griffiths (2009, 2010), who discuss a Bayesian iterated learning model (ILM) where each learner receives input from multiple members of the previous generation and has the task of inferring the distribution of languages within a population of agents. Using Bayesian learners sampling from the posterior distribution of hypotheses, their simulations show that convergence on the distribution determined by their prior biases depends on learners' expectations concerning the number of languages their input was produced from. These expectations are expressed in the concentration parameter α that parameterises the distribution over hypotheses. For larger values of α, learners expect more variety in their input, whereas for smaller values of α, they expect the data to have been produced by fewer languages. In the latter case, the outcome of the iterated learning process is determined by the input received by the first generation of learners, with learners sampling a hypothesis that magnifies the share of the language represented in the input more frequently. These findings are in line with those of Griffiths and Kalish (2005, 2007), who came to similar conclusions for transmission chains consisting of a single agent per generation.

Taking the results of Burkett and Griffiths (2009, 2010) as its starting point, this paper discusses a number of modifications and extensions to the original model. Burkett and Griffiths (2009, 2010) make a number of assumptions that, while not unreasonable, might affect the outcome of their simulations. Firstly, they assume that all agents within the population share the same value for α. Secondly, they assume that each item of input learners receive is produced by an agent randomly sampled from the population. Thirdly, learners only generate a single

hypothesis rather than learning incrementally, adapting their hypothesis in the light of each new data item. Finally, all transmission takes place vertically, with the learners' data being produced exclusively by agents from the previous generation. The first four sets of simulations presented here are intended to determine whether changing these assumptions affects the outcome of iterated learning. Secondly, as mentioned above, whether learners' prior biases or the input received by the first generation of learners play a greater role in determining the outcome of iterated learning depends on their expectations about how many languages their input was generated from; this is determined by the value of the concentration parameter α. Burkett and Griffiths (2009, 2010) leave open the matter of how α is assigned a value. One way of addressing this issue, namely learning both α and a hypothesis regarding the distribution of languages within the population in a hierarchical Bayesian model, is described and evaluated in the fifth and final set of simulations.

Chapter 2 provides the intellectual backdrop for the discussion to follow. The next two chapters provide descriptions of the frameworks used both here and by Burkett and Griffiths (2009, 2010), the iterated learning model and Bayesian learning. Chapter 5 presents the details of the two-language model used in Burkett and Griffiths (2009). Chapter 6 discusses some of the assumptions made by Burkett and Griffiths (2009, 2010) and presents the results of a series of simulations in each of which one of these assumptions is changed. Chapter 7 presents a model where agents learn a complex hypothesis consisting of both a distribution over languages and a value for α. The final chapter provides a general discussion of the findings presented here.

CHAPTER 2

Background

Estimates concerning the number of languages spoken in the world today vary widely (Crystal, 2000). Ethnologue (Lewis, 2009) gives the number of languages as 6909, 473 of which are classified as "nearly extinct"; many more have come and gone. Whatever the exact number, this degree of variety within the communication system of a single species is unique. Nonetheless, all languages can be analysed as having various levels of structure, be it phonological, syntactic or semantic, and a considerable amount of research is devoted to determining which *language universals*, i.e. underlying structural properties shared by all languages, exist.[1] Furthermore, all languages arguably share the property that they can be readily learned by human children with little or no explicit tutoring (pathologies notwithstanding). Indeed, it is conceivable that a child might be raised in a society other than that into which it was born and acquire the language of the former such that the language of the latter would leave no trace, and while there are ethical objections to testing this claim experimentally, research with Korean adoptees raised in France (Ventureyra et al., 2004) suggest that it is in fact correct.

Assuming language universals exist and children really are capable of acquiring any human language, it might be interesting to ask whether these two facts are in some way related and how they may be related. In other words, are children's language acquisition skills related to the structural similarities shared by all languages? If so, what is the nature of this relationship? The absence of any sizeable debate about the question suggests that the answer to the first question is considered to be "yes": although there is no *prima facie* reason to believe that this should

[1] That such universals *do* exist, while not wholly uncontroversial (see Evans and Levinson, 2009, for a somewhat polemical discussion of the subject), is widely assumed to be true.

necessarily be so, assuming it to be the case is without doubt a more parsimonious explanation than positing two separate mechanisms, one facilitating language acquisition, the other ensuring the existence of language universals.[2]

The second question, on the other hand, has been the source of considerable debate. Some researchers (e.g. Pinker and Bloom, 1990; Hauser et al., 2002; Jackendoff, 2003) have suggested that humans possess a biological endowment, commonly referred to as the *language faculty* (Fitch et al., 2005; Jackendoff and Pinker, 2005) or – "somewhat unfortunately", as Zuidema (2003) remarks – *Universal Grammar*. This *language organ* (Anderson and Lightfoot, 2002) may have emerged as a gradual adaptation, as suggested by Pinker and Bloom (1990) and Pinker and Jackendoff (2005), or as a spandrel, as Hauser et al. (2002) claim. It may consist of an ability to process recursive structures (Hauser et al., 2002; Fitch et al., 2005), a set of parameters describing linguistic features that are weighted statistically according to the input the language learner receives (Yang, 2004), or some rich set of features that evolved piecemeal (Jackendoff and Pinker, 2005). Regardless of its exact nature, the function of Universal Grammar is essentially always the same: it limits the number of possible grammars a child must consider when attempting to acquire a language. Without such constraints, it is claimed (Gold, 1967; Chomsky, 1965), children would be unable to determine the grammar of the language their are confronted with. Gold (1967)'s theoretical results are generally thought to support the argument from the poverty of the stimulus (POS), which claims that the input children receive when acquiring a language is insufficient, or too *impoverished*, for the language acquisition process to be successful unless some kind of constraints are in place to limit the number of possible grammars the learner will contemplate.[3] Since children clearly *are* capable of acquiring a language, there must be a set of constraints in place to enable this process. Furthermore, since these constraints cannot be learned, they must be innate and therefore biological. Since this set of constraints specifies which human languages are possible and therefore also determines what all languages have in common.

The "innateness hypothesis" (Putnam, 1967) and POS have been controversial since they were first proposed by Chomsky (1965). Nonetheless, over the past forty years, a great deal of research in linguistics has been based to a greater or lesser extent on some version or other of the hypothesis.[4] In recent years, however, it has come under increasing scrutiny, partly due

[2]Depending on how one is inclined to define what a language is, it may be all but impossible to posit two different mechanisms. See Steels (2000) for a brief discussion of the difficulties of defining a language.

[3]See Thomas (2002) for an overview of the history of POS.

[4]It is interesting to note that philosophers have been largely critical of the hypothesis. Putnam (1967), for example, states that the innateness hypothesis is "daring – or apparently daring; it may be meaningless, in which case it is not daring"; Goodman (1967) provides an amusing yet thorough critique of Chomsky (1965)'s ideas in the style of a Platonic dialogue. See also Mameli and Bateson (2006) and Godfrey-Smith (2007) for more recent

to growing doubts regarding the validity of POS. Firstly, a number of researchers (e.g. Pullum and Scholz, 2002; Sampson, 2005) suggest that the input children receive is not as impoverished as it is made out to be. Others (Scholz and Pullum, 2002; Zuidema, 2003) claim that Gold (1967)'s results do not support POS in the way many linguists believe them to. Some authors (e.g. Kirby, 2001; Zuidema, 2003; Chater and Christiansen, 2010) suggest that rather than being a problem that language learners must overcome, POS allows learners to modify the language they are learning, thereby making it more learnable. As Zuidema (2003) puts it, "the poverty of the stimulus *solves* the poverty of the stimulus (emphasis added)." On this view, languages are shaped to a considerable extent by the *process* of being transmitted from one generation to the next, and the transmission process is determined by biological *and* cultural factors that influence the learning process.

Secondly, although it is generally accepted that there must be some sort of constraints on the set of possible languages children will entertain when learning a language, it does not follow that these constraints must be innate or biological. That claim is based on the implicit assumption of a modularised mind where the *language* faculty is shielded in some way from developments in, say, the *social skills* faculty.[5] Some researchers (e.g. Goldberg, 2003; Tomasello, 2003, 2005; Sperber and Origgi, 2010) believe that constraints may arise from other aspects of a learner's maturational process. Tomasello (2003), for example, suggests that children are able to solve the *Gavagai* problem (Quine, 1975) because the joint attentional frame they construct when attending to something with mature speakers effectively restricts the set of possible meanings that they will entertain for unknown words to aspects of that attentional frame. The same self-imposed restriction may also allow children to determine cross-situational regularities regarding grammatical constructions (Goldberg, 2003). Similarly, Sperber and Origgi (2010) point out that pragmatic factors dependent on "naïve psychology" play an important role in language acquisition and evolution.[6]

Chater and Christiansen (2010) refer to the task language learners face as *C-induction*, or learning to coordinate with each other. The example they provide to distinguish this from *N-induction*, or learning to model the world, illustrates the distinction very well: imagine being

critical discussions of the concept of innateness.

[5] I am using the idea of a social skills faculty for the sake of argument; I am not suggesting that there is such a vague thing – although given that the vague idea of a language faculty is widely entertained, it is possible that someone will propose it.

[6] Such claims may be supported by analytic results concerning the population dynamics of language evolution (Nowak et al., 2000) which suggest that natural selection will only favour syntactic communication once there are a sufficiently large number of events where communication about these events contributes to individuals' fitness, that is, once there are enough things worth talking about. It could be argued that the human ability to use metarepresentations (Sperber, 2000) or to share intentionality (Tomasello, 2005) provides exactly that increase in the number of "newsworthy" events to initiate the emergence of syntactic communication while also restricting the possible meanings a hearer might entertain to exactly those metarepresentations or intentions. A detailed discussion of this idea is, unfortunately, outwith the scope of this dissertation.

asked how the sequence *1,2,3,...* continues. There are an infinite number of possible sequences that start in this way, *1,2,3,2,1,2,...*, say, or *1,2,3,1,2,3,1,...*, or even *1,2,3,π,-4.76,...*. However, Chater and Christiansen (2010) argue that the most natural way for us to *assume* that the sequence continues is *...4,5,6,...*. Since most people will continue the sequence in this fashion, they claim, this sequence will be suggested far more frquently than the number of possible continuations would lead one to expect. This is because C-induction is affected by cognitive biases shared by all humans: if everybody has a tendency to answer a question in the same way, that answer will come to dominate.[7] In other words, given incomplete information regarding a system (such as a sequence), humans will "fill in the gaps" on the basis of their own biases. Incorporating these biases into the system makes it easier for the system to remain intact through transmission. The idea of language as an adaptive system (e.g. Kirby, 2001, 2002; Kirby et al., 2008) relies heavily on the idea that, to ensure successful transmission, languages adapt to human biases, making it easier for learners to induce a language on the basis of a small subset thereof. This is possible precisely because the biases incorporated are shared by both teacher and learner.

In addition to arguing that the task facing learners is in fact easier than POS assumes it to be, it has also been suggested that learning from data is more powerful than granted by proponents of UG. Saffran et al. (1996) show that infants as young as 8 months are sensitive to the different transitional probabilities between syllables within nonce words and those across word boundaries after only a few minutes of exposure; Maye et al. (2002) found that even 6-month-olds are sensitive to distributional information regarding phonetic variation. Gerken et al. (2005) present evidence that, under certain circumstances, infants might be able to form syntactic categories on the basis of distributional cues. Research by Gómez (2002) and Gökaydin et al. (in press) shows that learners can employ different statistical analyses depending on the input they receive, and Monaghan et al. (2005) show that learners may use a combination of distributional and other cues when learning a language. A number of papers (e.g. Rowland et al., 2003; Theakston et al., 2003) suggest that the acquisition of syntactic constructions depends greatly on the type of input learners receive, and Perfors et al. (2010) show that child-directed speech allows Bayesian learners with domain-general abilities to infer a hierarchical phrase-structure grammar.

Together, these results cast doubt on the validity of POS. There are, however, other reasons to question the existence of an innate biological set of constraints on language acquisition.

[7] If one substitutes the term *paradigm* for *cognitive bias*, this is not unlike Kuhn (1962)'s view of science. Ironically, the UG paradigm can then be seen as exactly the kind of bias that leads researchers to answer questions in certain ways.

Christiansen and Chater (2008), for example, argue that the language faculty cannot be a biological adaptation because language change occurs too quickly for genetic adaptations for many linguistic features to evolve. Kirby et al. (2007) discuss how iterated learning with Bayesian learners selecting a hypothesis with maximum posterior probability can obscure the strength of the learners' biases, making it impossible for natural selection to act on strong biases, because what is being selected for, namely the outcome of the learning process, is not determined by the strength of the bias alone. Smith and Kirby (2008) present similar findings: cultural transmission can shield the language faculty from biological selection, making it less likely for a UG with strong constraints to evolve. Ladd et al. (2008) discuss an example of this, the differences between vowel formants in Italian and Yoruba: while these differences may, as Ladefoged (1984, cited by Ladd et al., 2008) suggests, be due to anatomical differences between the languages' speakers, these genetically inherited differences can be overcome: a Yoruba may still learn to speak perfect Italian and vice versa. Thus, although genetically determined traits have affected the languages spoken, the differences between the languages themselves have not arisen due to biological selection and can be overcome by learners. Finally, Dediu (2008) raises the question of whether models of language evolution relying on cultural processes (e.g. Kirby, 2001; Brighton, 2002) require a modern brain at all, and although the question is seemingly intended as a criticism of such models, the question can be turned around (as Tomasello, 2003, does): Why should one bother positing a biologically determined UG imposing strong constraints on language learners, when the cultural transmission process provides sufficient constraints to explain the emergence of human language and language universals?

Nobody denies that biological evolution plays *some* part in the emergence of language. The issue, rather, is whether biological selective pressures may have given rise to *domain-specific* constraints which in turn may have resulted in the phenomena outlined at the beginning of this chapter, viz., the possible existence of language universals and children's ability to acquire any human language. A growing body of evidence indicates that the constraints giving rise to language universals need not be determined biologically, but may arise from the language learning process itself. The task at hand, then, is to provide an alternative account of how language emerges. The following two chapters will discuss two frameworks that fufil different parts of this task: iterated learning and Bayesian learning.

CHAPTER 3

The Iterated Learning Model

The Iterated Learning Model (ILM) was introduced by Kirby (2001) as a way of modelling the effects on a language of being transmitted along chains of learners.[1] As a rule, it has four basic components (Kirby and Hurford, 2002):

1. one or more learning agents;
2. one or more teaching agents;
3. a meaning space; and
4. a signal space.

When implemented computationally, the first generation, or *cohort* (Smith, 2002), of agents is initialised with a set of data, commonly consisting of mappings between signals and meanings.[2] On the basis of these data, each agent then generates a hypothesis regarding the nature of these mappings. The agent then produces a set of such mappings itself, which is then used as input to the subsequent cohort in the transmission chain. Crucially, the set of mappings produced by the agent also includes mappings for meanings it did not encounter in the data it learned from. The agent receives no positive or negative feedback during this phase, making the outcome of

[1] Note that this is not meant to imply that the ILM refers exclusively to chains of single learners; a scenario where multiple learners have multiple, possibly identical, teachers may still be implemented as an ILM. Niyogi and Berwick (2009)'s criticism of the ILM in this particular respect suggests an exceptionally narrow interpretation of the model.

[2] This need not necessarily be the case; whether the specification of distinct meaning and signal spaces are required depends on what is being modelled. The models reported in later chapters, for example, do not distinguish between a meaning and a signal space because the distinction is not relevant to what is being modelled. Nevertheless, a large number of ILMS *do* make use of separate meaning and signal spaces.

IL independent of an agent's communicative success. The initialisation step aside, this process is repeated a large number of times, or until a stable state is attained where successive cohorts produce the same mappings.

The ILM has proved very popular, with numerous *in silico* experiments being conducted within the framework. Many of these simulations (Hurford, 2000; Kirby, 2001; Brighton, 2002; Smith et al., 2003) suggest that the emergence of languages with features such as generality or regularity – universals whose emergence has been modelled using ILMs (Kirby, 2001, e.g.) – can emerge from the process by which languages are transmitted rather than any built-in properties of the agents involved. The main factors determining whether or not compositional languages emerge are

- the amount of data available to the learner relative to the size of the language being learnt, widely referred to as the *transmission bottleneck* (Kirby, 2001; Brighton, 2002); and
- the structure of the meaning space.

The first point is obvious at the limit: if an agent were provided with a data set containing a mapping for each meaning, it would be able simply to reproduce the appropriate signal when called upon to produce a signal itself. The language system would be a static one. As Brighton (2002) points out, in the ILM, it is the transmission bottleneck that causes POS. The interesting point is that in the ILM, POS affects the structure of the language agents acquire. Mappings frequently present in the data are more faithfully reproduced than infrequent ones; as a result, the latter are more likely to be subject to regularisation when agents are called upon to produce signals for them (Kirby, 2001; Hurford, 2000).

The second point, on the other hand, is slightly more difficult to deal with. One point criticised, for example, by Swarup and Gasser (2009) and Smith (2001, 2005) is that, in many implementations of the ILM, agents share a common meaning space – unlike humans, who are generally thought to have minds of their own. However, Kirby (2007) and Smith (2005) show that the outcome of IL is essentially the same if learners do not share a common meaning space. Moreover, it has been shown (e.g. Steels, 2000; Smith, 2001; Barr, 2004) that agents are capable of constructing individual meaning spaces and using them to communicate. Thus, while many implementations of the ILM provide their agents with a common meaning space, this is can be seen as an implementational simplification rather than a prerequisite for the success of such models.

Both Kalish et al. (2007) and Mesoudi et al. (2006) point out that iterated learning experiments with human subjects were being conducted as early as the 1930s (Bartlett, 1932, referred

to in Mesoudi et al., 2006), albeit not to test hypotheses about language. Recently there have been a number of new experiments implementing ILMs with humans rather than software agents (e.g. Kalish et al., 2007; Kirby et al., 2008; Beppu and Griffiths, 2009; Xu et al., 2010, Perfors and Navarro, in press) for the purpose of validating theoretical findings obtained from computational experiments.[3] Kirby et al. (2008), for example, were able to confirm that iterated learning in chains of humans resulted in linguistic structure emerging, and Perfors and Navarro (in press) showed that the outcome of iterated learning with humans depends, *inter alia*, on the structure of the meaning space. However, more experimental work is required to determine which theoretical results from simulations within the IL framework apply to human learners.

The limited amount of experimental replication of results is not the only criticism that can be levelled at the ILM. The model itself is not without flaws, and critics have been quick to point them out. These flaws are

1. the absence of horizontal transmission, i.e. learning from peers (Vogt, 2005);

2. the use of transmission chains consisting of a single learner per generation (Niyogi and Berwick, 2009; Dediu, 2009);

3. the possibility of the emergent universal such as compositionality being built in to the model (Swarup and Gasser, 2009).[4]

Each of these criticisms, however, has been addressed. For example, Vogt (2005) showed that allowing horizontal as well as vertical transmission acted as an *implicit* bottleneck, rendering an explicit bottleneck imposed by the modeller superfluous, and Niyogi and Berwick (2009)'s social learning model is an ILM that allows learners to receive input from multiple adults, as is Smith (2009)'s model.[5]

As for compositionality being built in to the model, there are, I believe, two ways in which this may be the case. Firstly, by providing a structured meaning space to communicate about, modellers might be providing a "blueprint" of sorts for structured communication about those meanings. This may be inevitable (or even desirable) if one assumes that linguistic structure reflects meaning structure in some way (Kirby, 2007). Secondly, the algorithm learners use to process the input they receive may in some way contain the basis of linguistic structure which emerges as a result of IL. The first case has been dealt with, as mentioned previously, in

[3] See Mesoudi and Whiten (2008) for further examples, as well as a review of alternative methods for modelling cultural transmission, such as the replacement method implemented by Smith (2002).

[4] Compositionality is mentioned here because it is the language univesal which has attracted the most attention with regard to the ILM (Kirby et al., 2007).

[5] Niyogi and Berwick (2009)'s point about chains of single learners possibly influencing the dynamics of learning is certainly valid, the suggestion that social learning is something other than a type of ILM less so.

experiments concerned with the construction of a meaning space (e.g. Smith, 2001; Steels, 2000; Kirby, 2007).[6] If one assumes that agents have the ability to structure their meaning space in some fashion – an assumption that is not unreasonable, given that humans appear to have such an ability (Kemp et al., 2007) –, then it is possible for agents to communicate about the objects associated with the meanings they construct. The second issue is due at least in part to the variety of algorithms that agents are implemented with in ILMs: while the model provides a clear description of the transmission process, the issue of how to model agents' cognitive processes goes largely unanswered. In Hurford (2000) and Kirby (2001), for example, agents use heuristic algorithms to induce a grammar, whereas Smith et al. (2003) implements agents as associative networks.[7] This variety makes it more difficult to compare different models and might give the impression that agents are simply implemented in such a way that they cannot help but generate compositional languages. Moreover, there is often little in the way of psychological motivation for the various implementations. Fortunately, Bayesian learning can provide just such a psychologically motivated model for implementing agents, while also going some way to alleviate the criticism of built-in compositionality.

[6] Note that this issue is closely related to the problem of grounding symbols in the world; see Vogt (2002); Vogt and Divina (2007) for further discussion.

[7] I am by no means singling out these papers; I have chosen them for purely expository purposes.

CHAPTER 4

Bayesian learning

As mentioned in chapter 3, the manner in which agents have been implemented in ILMs raises a number of issues: the lack of psychological motivation, the risk of building linguistic universals into the model, and the difficulty comparing results from different models. Bayesian learning provides a way of addressing all three of these issues.

A number of researchers (e.g. Chater and Manning, 2006; Chater and Christiansen, 2010; Griffiths, 2011; Tenenbaum et al., 2011) have noted that learning a language is a problem of induction or inference: on the basis of limited data, learners attempt to determine which utterances are permissible in a language. That much is, I believe, fairly uncontroversial. Probability theory provides a way of formalising questions of inference in a precise manner. More specifically, Bayesian probability theory has been used to formalise the kind of task language learners face: given a set of data d and a *prior (inductive) bias*, i.e. a preference for a particular solution to the task at hand, learners can infer a solution, or *hypothesis*, by employing *Bayes' rule* (or *Bayes' theorem*):

$$p(h|d) = \frac{p(d|h)p(h)}{p(d)}, \tag{4.1}$$

where $p(h)$ is the *prior probability* – that is, the inductive bias learners bring to the learning task – of the hypothesis h, $p(d|h)$ is the *likelihood* of the data d being generated by h, and $p(d)$ is the overall probability of d, which can be determined by marginalization:

$$p(d) = \sum_{h' \in \mathcal{H}} p(d|h')p(h'), \tag{4.2}$$

where $\mathcal{H}$ is the *hypothesis space*, the set of all hypotheses the learner chooses from. Since the value of $p(d)$ is constant for a fixed hypothesis space, equation 4.1 can be rewritten as

$$p(h|d) \propto p(d|h)p(h). \tag{4.3}$$

Given these formulae, a learner can calculate a distribution over hypotheses and then select a hypothesis from this distribution. One way of doing so is simply to choose the hypothesis with the highest posterior, or *maximum a posteriori*, distribution. An alternative method is to sample a hypothesis when required according to their posterior distribution: thus a hypothesis with twice the posterior probability of another hypothesis ought on average to be sampled twice as often. How learners select hypotheses can have a profound impact on the dynamics of a model.

Although Bayes' rule has been known for many years – Bayes' work was originally published in 1763 –, Bayesian modelling of cognitive processes has recently become very popular, so much so that Shultz (2007) speaks of "a revolution in cognitive science". In recent years, Bayesian inference has been used, among other things, to model category learning (Kemp et al., 2007), reasoning about objects (Téglás et al., 2011), word learning (Xu and Tenenbaum, 2000, 2007; Frank et al., 2009; Xu et al., 2010) and language learning and evolution (Griffiths and Kalish, 2005, 2007; Kirby et al., 2007; Smith and Kirby, 2008; Smith, 2009; Burkett and Griffiths, 2009, 2010; Perfors et al., 2010; Perfors and Navarro, ress). One advantage of modelling agents as Bayesian learners is that the inference algorithm is determined not by the modeller – one of the criticisms raised in the previous chapter – but by probability theory; this makes it easier to compare results. It also provides a psychological basis for the cognitive process agents are provided with, bringing theoretical work with ILMs closer in line with experimental research in psychology.[1] Another advantage is that a considerable amount of mathematical research has been carried out in the field of probability theory. These results can be used in theoretical models of cognition and then tested experimentally, to determine the best way of modelling cognitive processes using probability theory.

There are, however, a number of possible objections to modelling humans as Bayesian learners. Firstly, we have no knowledge of how humans might carry out such inferential calculations. This objection can be countered relatively easily. Marr (1982) distinguishes three levels of analysis with regard to cognitive processes. The *computational* level of analysis is concerned with determining the task a cognitive process is meant to accomplish and the ideal way of doing so.

[1] This psychological foundation makes it more justifiable to refer to ILMs as *agent-based cognitive models* (Vogt, 2009).

For example, to say that learning a language is about learning a language provides little insight; to say that learning a language is about inferring a grammar on the basis of sparse data, on the other hand, specifies the task in a way that makes it easier to determine the "ingredients" necessary to accomplish it. The *algorithmic* level deals with the algorithm used to accomplish the task at hand and the representations needed to do so. Finally, the *implementational* level is concerned with how the algorithm used for a task is physically implemented, e.g. in a human's brain. The three levels differ in the degree of abstraction with which they analyse a task, with the computational level providing the most abstract view. It is at this level that most research using Bayesian learning is conducted. The objection voiced above, on the other hand, refers to the implementational level. Clearly more research is needed to determine how Bayesian models are implemented and which algorithms are used to infer hypotheses (some such work is underway; see, for example, Sanborn and Griffiths, 2006; Fiser et al., 2010); however, the fact that we know little about the implementational details of Bayesian learning does not count against a Bayesian analysis of cognition at the computational level.

Secondly, humans are not particularly good at calculating probabilities. Then again, most humans are also not particularly good at carrying out differentiation in their heads, either – yet they are perfectly capable of determining the trajectory of a ball so as to catch it. Similarly, even with limited metalinguistic skills, people are capable of using language (although see Dabrowska, 2006, on the relationship between metalinguistic knowledge and grammaticality judgements). Thus the fact that humans appear largely unable to consciously carry out a task they are meant to be carrying out constantly need not count against Bayesian learning.

Thirdly, and more seriously, there is the question of how to determine the hypothesis space for a particular task. If they are innate, then Bayesian language learning simply provides an additional account of how UG might work. One possible solution is to say that the hypothesis space is determined by the possible computations that the underlying implementation can carry out. If that is so, then the hypothesis space is simply given *a priori*. Perfors et al. (2011) propose distinguishing between the *latent hypothesis space* that contains all potential hypotheses and the *explicit hypothesis space* that contains the hypotheses actually considered by a learner; however, they do not explain how a hypothesis is generated, i.e. transferred from the latent to the explicit hypothesis space.

Finally, a similar question arises concerning the prior bias. Must these biases be innate? No. Griffiths and Kalish (2007) suggest that prior biases are simply a convenient way of summarising all of the factors that might influence how a learner entertains. It may be that the outcome of one learning process results in a prior bias regarding some other task. But this explanation can only

go so far without leading to an infinite regress: the explanation of where prior biases originate cannot be "it's priors all the way down", to paraphrase Hawking (1988). Nevertheless, it may be possible that the prior biases for most tasks are derived from a small number of innate *a priori* biases. However, which prior biases might depend on which and what these dependencies look like is not an issue that can be dealt with on the computational level of analysis alone.

4.1 Bayesian language learning

Bayesian models have been used to model the evolution of language in a number of papers (see references above), providing new and occasionally surprising insights into the factors that shape language as it emerges by cultural transmission. One of these surprising results is that of Griffiths and Kalish (2005, 2007), who show that the outcome of iterated learning by agents who are Bayesian learners that choose a language by sampling from the posterior distribution over possible languages simply reflects the agents' prior distribution over those languages. This finding is analytic, with the iterative process being modelled as a Markov chain. Unlike simulations with agents implemented using other algorithms, the transmission bottleneck does not affect the actual result of IL, merely the speed at which convergence occurs.

Kirby et al. (2007) expand on this result, demonstrating that Bayesian learners selecting a hypothesis with maximum posterior probability *are* affected by the bottleneck, since the latter affects the distribution over hypotheses an agent entertains by amplifying weak prior biases. These findings suggest that it is not only the prior bias (and hence the hypothesis space), but also agents' strategies for selecting a hypothesis from the posterior, that affects the outcome of Bayesian iterated learning. They are supported by Smith and Kirby (2008), who show that these results hold under a number of different evolutionary scenarios. Cultural transmission can shield weak biases from biological selection, since fitness – which is what ought to result in selective pressures – is determined not just by the (possibly hereditary) prior bias, but also by the (cultural) transmission process. Languages may, then, be shaped by the process of IL and cultural transmission, if language learners are Bayesian learners selecting a hypothesis with maximum posterior probability.

Dediu (2008), however, raises an issue concerning the results of both Griffiths and Kalish (2007) and Kirby et al. (2007), an issue relating back to the second criticism of the ILM mentioned above. He shows that increasing the number of agents per cohort in each transmission chain does not greatly affect the outcome of IL *if* both agents employ the same hypothesis selection strategy. However, if agents in paired transmission chains use *different* strategies, the

outcome more closely resembles that of transmission chains consisting only of agents sampling from the posterior. In other words, it is not permissible simply to assume that results obtained from transmission chains of single agents also hold for more complex chains – a point harking back to Niyogi and Berwick (2009). Smith (2009) makes a similar point: if sampling Bayesian learners are in fact facing the task of choosing a hypothesis on the basis of data generated by multiple agents, the outcome of IL will not simply converge on learners' prior biases. Instead, transmission factors like the starting conditions of the simulation, the amount of data (i.e. the transmission bottleneck that Griffiths and Kalish (2005, 2007) showed had no impact on the ultimate outcome of IL) and an interaction between the amount of data and the prior, affect the posterior distribution. Ferdinand and Zuidema (2009) reach a similar conclusion; however, their analysis adds an important insight. If agents are sampling from the posterior distribution, the agents within a population at time t will not necessarily all choose the same hypothesis with which to produce data for the following cohort. This means that the data will be a sample from the product of the agents' individual posterior distributions. If learners are simply considering the probability of the data being generated by a single distribution over hypotheses, the distribution they are attempting to determine, namely the one that generated the data, may not be among the distributions they are evaluating. Once this is the case, Ferdinand and Zuidema (2009) argue, the agents are no longer behaving like rational Bayesian learners.

CHAPTER 5

The Burkett and Griffiths (2009) model

Burkett and Griffiths (2009, 2010)[1] take Ferdinand and Zuidema (2009)'s insight and modify the scenario outlined in Smith (2009). There, a population of agents consisting of discrete cohorts, adults and learners, learn a single language by sampling a hypothesis regarding the language spoken on the basis of data produced by agents in the previous cohort who themselves learned a language from data provided by the previous cohort, in accordance with the ILM.[2] However, they learn only one language, despite the adult population generating data on the basis of two different languages. Burkett and Griffiths (2009) suggest that such learners are not behaving like rational Bayesian agents; instead, learners ought to take into account that the data are produced by different speakers who might also be speaking different languages, that is, they should be attempting to determine the *distribution of languages within the population.* Once they do so, the probability distribution over the data learners receive is no longer[3]

$$p(d) = \sum_{h \in \mathcal{H}} p(h)(\prod_{w \in d} p(w|h)). \tag{5.1}$$

Here, it is assumed that all words are produced by the same hypothesis, and $p(d)$ can be calculated by summing over product of the likelihoods of each hypothesis producing these data

[1] For the remainder of the paper, I shall refer mostly to Burkett and Griffiths (2009), since both papers essentially report the same findings, but Burkett and Griffiths (2009) also report the results of simulations using a two-language scenario like the one used here.

[2] The data for the first cohort of learners is generated by sampling from the base distribution G_0, described below.

[3] Most of the formulae in this section are taken from Burkett and Griffiths (2009).

and the priors of the hypotheses. Instead, the distribution describing the data d is

$$p(d) = \prod_{w \in d} (\sum_{h \in \mathcal{H}} p(h)p(w|h)), \tag{5.2}$$

where for each word $w \in d$, one must first determine the likelihood of it being produced by each hypothesis $h \in \mathcal{H}$ before determining p(d) as the product of the likelihood of each word under all hypotheses.

Burkett and Griffiths (2009) discuss two ways in which rational Bayesian agents might take into account the diversity of hypotheses generating the data they receive; I shall only discuss the second possibility, that of learning a *distribution over languages* rather than a particular language. Furthermore, I shall discuss only the simplest case, where agents must learn a distribution over just two languages l_0 and l_1, each consisting of a single word, w_0 and w_1 respectively. Since the hypothesis space $\mathcal{H}$ that learners are operating in no longer consists of languages but distributions over languages, and there are an arbitrary (and potentially infinite) number of possible distributions over two languages for learners to choose from, Burkett and Griffiths (2009) propose modelling the hypothesis using a Dirichlet process (DP).[4] The prior of a DP is parameterised by the *base distribution* G_0, which specifies the learner's *a priori* set of preferences for the languages in $\mathcal{H}$, and the *concentration parameter* α, which affects the number of languages a learner believes there to be: the greater the value of α, the smaller the number of languages a learner will expect *a priori*. G_0, in turn, is parameterised by p_0 such that

$$p_{G_0}(l_i) = p_0^{\delta_{i,0}}(1 - p_0)^{\delta_{i,1}}, \tag{5.3}$$

where $\delta_{i,j}$ is Kronecker's delta. In other words, the probability of l_0 is p_0, that of l_1 is $(1 - p_0)$.

In the two-language scenario, equation 5.2 can be expanded to describe the probability distribution of the data learners receive as

$$p(d) = \prod_{w \in d} (\sum_{h} p(h) \sum_{l} p(l|h)p(w|l)), \tag{5.4}$$

with the probability of producing a word w_i given by the formula

$$p(w_i|l_j) = (1 - \epsilon)^{\delta_{i,j}} \epsilon^{\delta_{(1-i),j}}, \tag{5.5}$$

where $\delta_{i,j}$ is again Kronecker's delta and ϵ is a suitably small value representing the probability

[4] See Frigyik et al. (2010) for a thorough introduction to Dirichlet processes.

of noise during production. Thus a speaker of l_0 will produce w_0 with probability $(1-\epsilon)$ and w_1 with probability ϵ. Each learner a in the learner cohort ($a \in A$) receives $|d| = 20$ data items, each of which is generated by first uniformly sampling an agent from the adult cohort that then samples a language according to its hypothesis before finally producing a word from the language selected according to equation 5.5.

To determine a hypothesis, learners in the Burkett and Griffiths (2009) model use a Gibbs sampler implementing a Chinese Restaurant Process[5]: first, each $w \in d$ is assigned to a cluster c_w, and each cluster c is assigned a language l_c. Which cluster c_w is sampled for a particular word is determined by the distribution

$$p(c|w, l_c, \mathcal{C}) \propto \begin{cases} n_c p(w|l_c) & c \text{ is an existing cluster} \\ \alpha \sum_l G_0(l) p(w|l) & c \text{ is a new cluster} \end{cases}, \tag{5.6}$$

where $\mathcal{C}$ is the set of all current clusters and n_c is the number of words assigned to a cluster c, defined as

$$n_c = \sum_{w \in d} \delta_{c, c_w}. \tag{5.7}$$

This "rich get richer" approach means that clusters with more words assigned to them are more likely to get a new word assigned to them.

The language l_c assigned to a cluster is determined by sampling from the distribution

$$p(l|w, c) \propto G_0(l) \prod_{w \in d} p(w|l)^{\delta_{c_w, c}}. \tag{5.8}$$

A hypothesis h is defined by

$$p(l|h) = \frac{n_l + \alpha G_0(l)}{|d| + \alpha}, \tag{5.9}$$

where n_l is the number of clusters assigned the language l:

$$n_l = \sum_c n_c \delta_{l, l_c} \tag{5.10}$$

It is clear from equation 5.9 that the value of α determines the extent to which G_0 affects a learner's hypothesis. This sampling process is repeated for a number of times (five times in Burkett and Griffiths, 2009), the final iteration providing the learner with a hypothesis regarding the distribution of languages within the population.[6]

Burkett and Griffiths (2009) ran their simulations with 100 agents per cohort ($|A| = 100$),

[5] See Navarro and Perfors, 2010 for an excellent introduction to the Chinese Restaurant Process.

[6] The earlier iterations are used to "burn in" the Gibbs sampler.

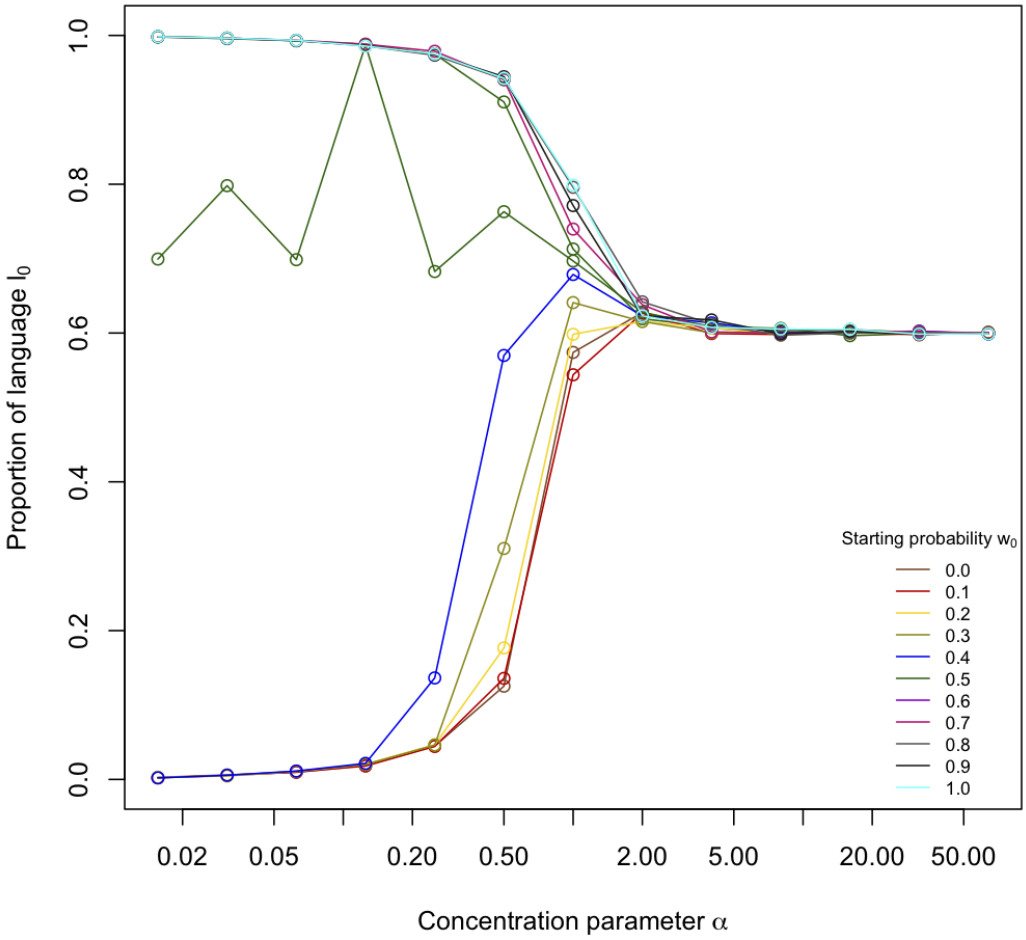

Figure 5.1: Replication of Burkett and Griffiths (2009)'s results of learning the distribution over two languages with a Dirichlet process prior.

each of which received 20 data items from which to learn ($|d| = 20$). Each simulation ran for 50 generations with $p_0 = 0.6$ and $\epsilon = 0.05$. Figure 5.1 shows a replication of their results, with the proportion of l_0 speakers averaged over ten simulation runs (as opposed to 200 runs in the original paper).[7] The different graphs represent simulations with different initial probabilities for the production of w_0. The graphs closely match those in Burkett and Griffiths (2009) except that for $p(w_0) = 0.5$. However, according to K. Smith (personal communication, August 2011), a high degree of variation in that particular case is relatively common in replications of these results. Regardless of the structure of the initial data received by the first generation of learners, if α is larger than approximately 8, the posterior distribution over hypotheses clearly converges on the prior G_0 defined in equation 5.3. As the value of α decreases, however, the outcome of iterated learning increasingly magnifies weak biases in favour of l_0 or l_1. For example, even with a weak bias in favour of l_0, e.g. when $p(w_0) = 0.6$, that language dominates the population

[7]The source code used in this dissertation is available here.

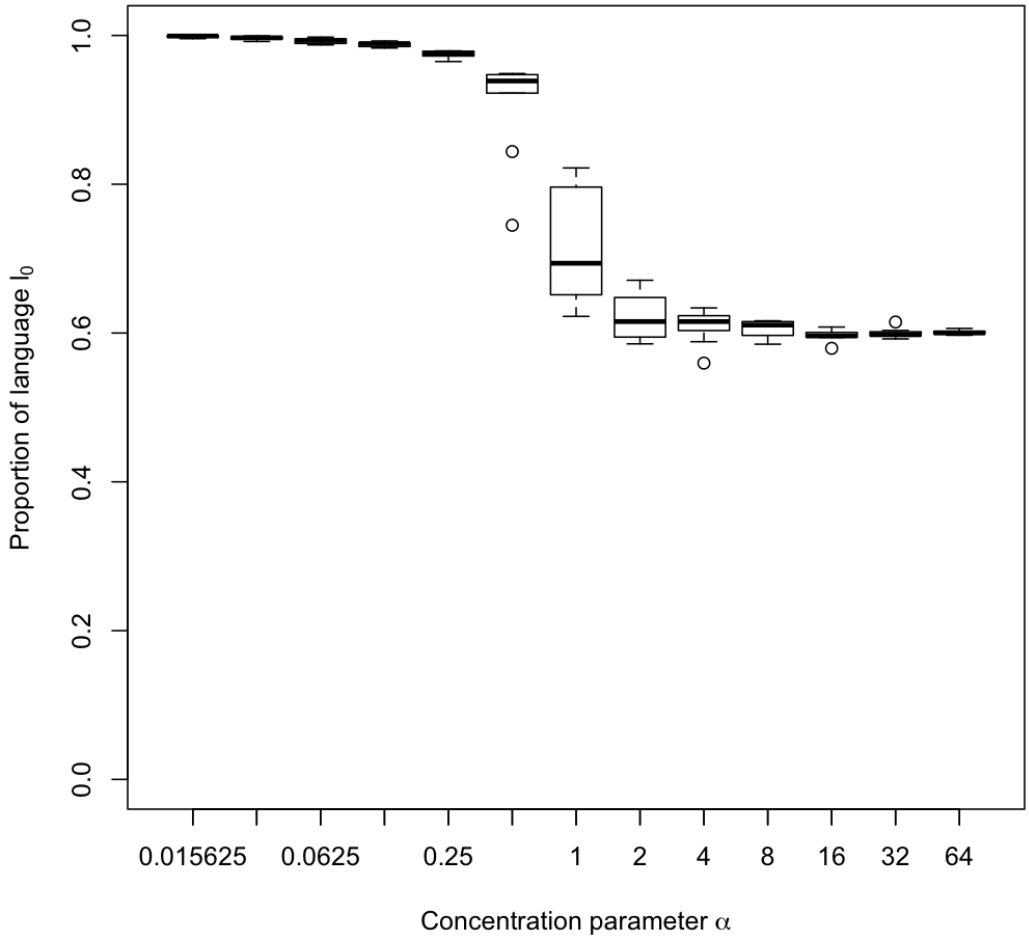

Figure 5.2: Proportion of l_0 spoken for different values of α, with the starting probability of $w_0 = 0.6$. Note the different scale of the x axis compared to figure 5.1.

for $\alpha \leq 0.5$. The transition from convergence to the prior distribution over hypotheses to magnification of potentially weak prior biases takes place within the interval $0.125 \leq \alpha \leq 8$. As figure 5.2 shows, populations with α values within this interval show greater variance with regard to the proportion of l_0 spoken. Variance is particularly great when $\alpha = 1$.

5.1 A brief justification

Before continuing, it might be worthwhile briefly to ask whether such a model, where learners choose between two very simple languages, but do so repeatedly by sampling from a distribution over languages, can actually provide any insight into what humans are actually doing when learning a language. Surely, one might object, even bilinguals don't go around deciding whether to make an utterance in, say, German or English!

It is indeed the case that people needn't often choose between two languages without knowing

which one would be more appropriate.[8] If one thinks of what are being called languages as alternatives *within* a language, the situation becomes more plausible. Take, for example, the utterances *kick down the door* and *kick the door down*. Both are correct and mean the same thing, yet they are based on different constructions. If l_0 and l_1 are interpreted as abstractions of constructions of this kind, one can imagine learners acquiring and using them both, choosing them according to their hypothesis regarding the distribution of the constructions in the English language.

I hope this brief example provides sufficient justification to continue exploring the simple two-language model.

5.2 Conclusion

The replication of Burkett and Griffiths (2009)'s results for the two-language scenario demonstrates that the implementation used here approximates that of Burkett and Griffiths (2009) sufficiently to carry out simulations that modify some of the assumptions their model makes.

[8] An exception to this rule does spring to mind, however. In the border region between (German-speaking) Switzerland and (francophone) France, it is common courtesy for hikers to greet one another. How should one greet strangers, "Grüezi" or "bonjour", if one wishes to greet them in their native language? One might generate a hypothesis over greetings on the basis of previous encounters, i.e. $p(d)$, and one's general inclination to use a foreign language, i.e. one's prior bias.

CHAPTER 6

Modifications to the Burkett and Griffiths (2009) model

6.1 Introduction

The Burkett and Griffiths (2009) model makes a number of assumptions:

1. All agents within the population have the same value for α.
2. Each word in a learner's input is sampled uniformly from the entire population.
3. Each word in a learner's input is generated by an agent from the preceding cohort.
4. Learners generate a single hypothesis on the basis of their entire input.

Each of these assumptions help to simplify the model; however, they also make the model less realistic. This is not a fault *per se*: the more realistic a model is, the more complex it becomes, making it more likely that confounding factors may be introduced unwittingly. Nevertheless, it is not unreasonable to make changes to the model to see whether the results hold in slightly more realistic scenarios. The following simulations each modify a single assumption from the list above.

6.2 Modification 1: heterogeneous population

Although not stated explicitly, it is clear from their exposition that in Burkett and Griffiths (2009)'s model, all agents share the same value for α. While this assumption may not be unreasonable, as it stands it appears to be a simplification for the purpose of modelling: regardless

of how α might be assigned a value, there is no *prima facie* reason to take for granted that a population of learners will be homogeneous in this respect. If the value is assigned by some hereditary mechanism or other, a degree of variation would be required for natural selection to choose a value from. If, on the other hand, the value of α is acquired, it may be influenced by other factors known to affect language skills, such as a learner's socioeconomic status (Hoff, 2003) or metalinguistic knowledge (Dabrowska, 2006), e.g. concerning the number of languages being spoken. If a learner's concentration parameter is itself set by sampling a distribution arrived at by Bayesian learning, one would again expect the population to be heterogeneous with regards to the value of α.[1]

In the first set of simulations, the assumption of a homogeneous population with regard to the value of α is dropped. Instead, each cohort consists of two subsets of agents, A_1 and A_2, with $\alpha_1 = 32$ and $\alpha_2 = 0.125$. These values were chosen because in homogeneous populations, the two result in very different proportions of l_0 being spoken (see Figure 5.1). Four different simulations were run, with $|A_1| \in \{50, 33, 25, 20\}$, respectively, and $|A_2| = |A| - |A_1|$. This change apart, the simulations were run under the same conditions as the Burkett and Griffiths (2009) model, i.e. $|A| = 100$, $|d| = 20$, $p_0 = 0.6$ and $p(w_0) = 0.6$.[2]

6.2.1 Results

Figure 6.1 shows the results of running simulations with different proportions of agents with $\alpha = 32$ and $\alpha = 0.125$, respectively. It is clear from the graphs that the agents' hypotheses become more skewed in favour of l_0 as the share of agents with $\alpha = 0.125$ increases. Furthermore, the results show far greater variance than in the original Burkett and Griffiths (2009) model. However, the proportion of l_0 spoken is not simply a reflection of the population's mean α value. Take, for example, the case where $\alpha = 32$ for 25% of the agents within each cohort. This is equivalent to a mean α of approximately 8 for the entire population.[3] As figure 5.1 shows, in a homogeneous population, the proportion of l_0 spoken when $\alpha = 8$ is approximately that of the prior, i.e. 0.6. In a heterogeneous population, however, this proportion increases to over 0.8. Figure 6.2 makes the difference clearer by comparing the outcome of IL for homogeneous populations with α values of 8 and 16, respectively, with IL in heterogeneous populations whose means for α approximate these values.[4]

At first sight, it seems as though there were additional dynamics at work affecting the agents'

[1] This point is reminiscent of the one raised by Ferdinand and Zuidema (2009) regarding the Bayesian rationality of populations of Bayesian learners sampling hypotheses from the posterior.

[2] Unless stated otherwise, these values were used for all simulations reported in this chapter; all results are averaged over 10 runs.

[3] Specifically, $0.25 * 32 + 0.75 * 0.125 = 8.09375$.

[4] This is the case where the proportion of agents with $\alpha = 32$ is 25% and 50%, respectively.

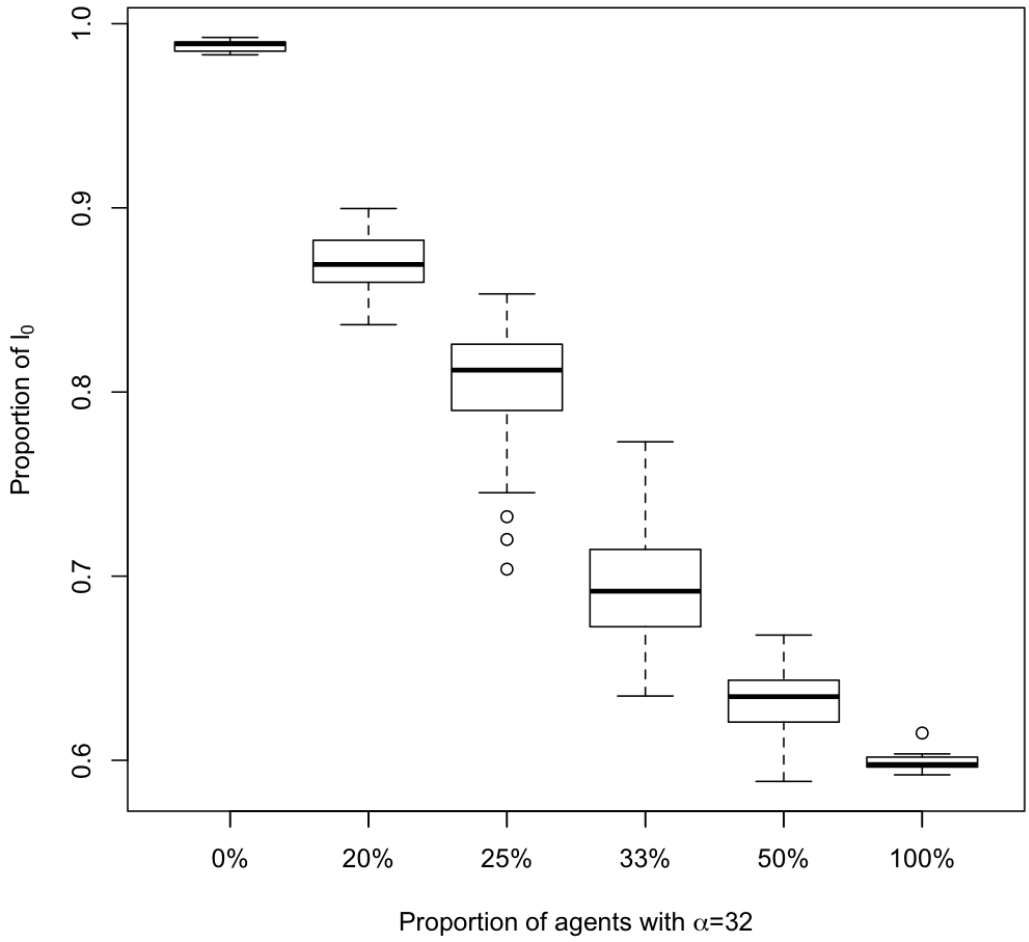

Figure 6.1: Comparison of simulations with populations consisting of different proportions of agents with $\alpha = 32$ and $\alpha = 0.125$. Note the different scale of the y axis compared to figure 5.2.

hypotheses. The effect of this is reminiscent of that reported in Smith (2009): an increase in the proportion of learners with $\alpha = 0.125$ has a disproportionate effect on the outcome of iterated learning. On reflection, however, an alternative explanation building on Ferdinand and Zuidema (2009)'s argument seems more reasonable. Agents in this simulation work on the assumption that the data they receive was generated by agents with the same value for α as they have. Clearly that will not be case: when $\alpha = 32$ for, say, 50% of the population, on average, half of an agent's data will have been produced by an agent with an α value different from its own. The hypothesis that produced a learner's data will therefore more often than not be outwith the learner's hypothesis space: the learner is no longer a rational Bayesian agent. To regain their Bayesian rationality, agents would have to take into account the fact that different words were potentially produced by agents with different α values. One way of dealing with this is discussed in chapter 7.

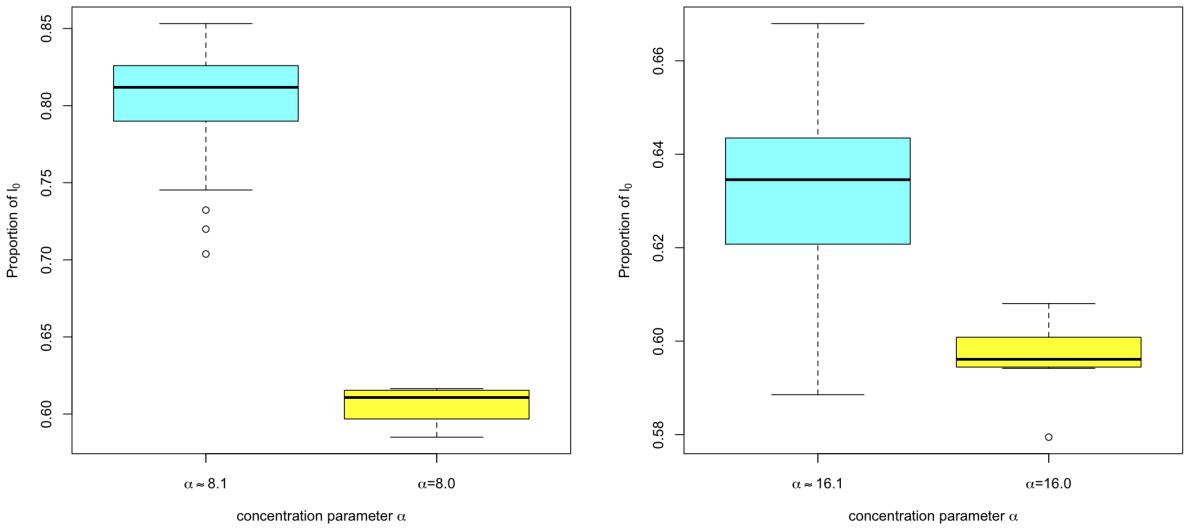

Figure 6.2: Comparison of the outcome of iterated learning in heterogeneous populations (blue), shown here with approximate average α values, and homogeneous populations (yellow) with a similar average α value.

6.3 Modification 2: cultural parents

In Burkett and Griffiths (2009), the agents producing input for each learner are drawn uniformly from the preceding cohort. As a result, each word a learner hears might be uttered by a different agent. However, this is not normally the case for human language learners. On the one hand, language communities may be too large or too widely dispersed to allow teachers to be drawn so uniformly. On the other hand, most children will be exposed more to the language of their principle and secondary caregivers – such as their parents, other relatives, and their parents' friends – than to that of the proverbial man on the Clapham omnibus. This has implications for the variety of input children receive. For example, the input caregivers provide their children with is determined in part by their socioeconomic status (Hoff, 2003). Furthermore, children do not appear to weight the contribution of different caregivers equally (Pancsofar and Vernon-Feagans, 2006). Thus, even if one assumes that learners receive equal amounts of input with which to determine a posterior distribution over hypotheses, the input may have been generated from small subsets of the speaker community that may differ greatly for different learners.

To approximate this situation, the data sampling process was modified. Rather than sampling an adult for each data item to be generated, a fixed number n_p of *cultural parents* were sampled from the adult cohort. Each cultural parent then generated the same number $|d_p|$ of data items. In each simulation, n_p and $|d_p|$ were chosen such that $|d| = n_p|d_p| = 20$. This sampling procedure increases the likelihood of an agent a_i being selected to generate $|d_p|$ items from $(\frac{1}{|A|})^{|d_p|}$ to $\frac{1}{|A|}$ and ought on average to lead to a reduction in the number of hypotheses used to generate the data a learner receives while maintaining the same total amount of input. It not

only more closely resembles the situation faced by human learners, but also makes it possible to separate the effects of reducing variety in learners' input from the amount of input they receive. Merely reducing the amount of data learners receive would confound these two factors.

6.3.1 Results

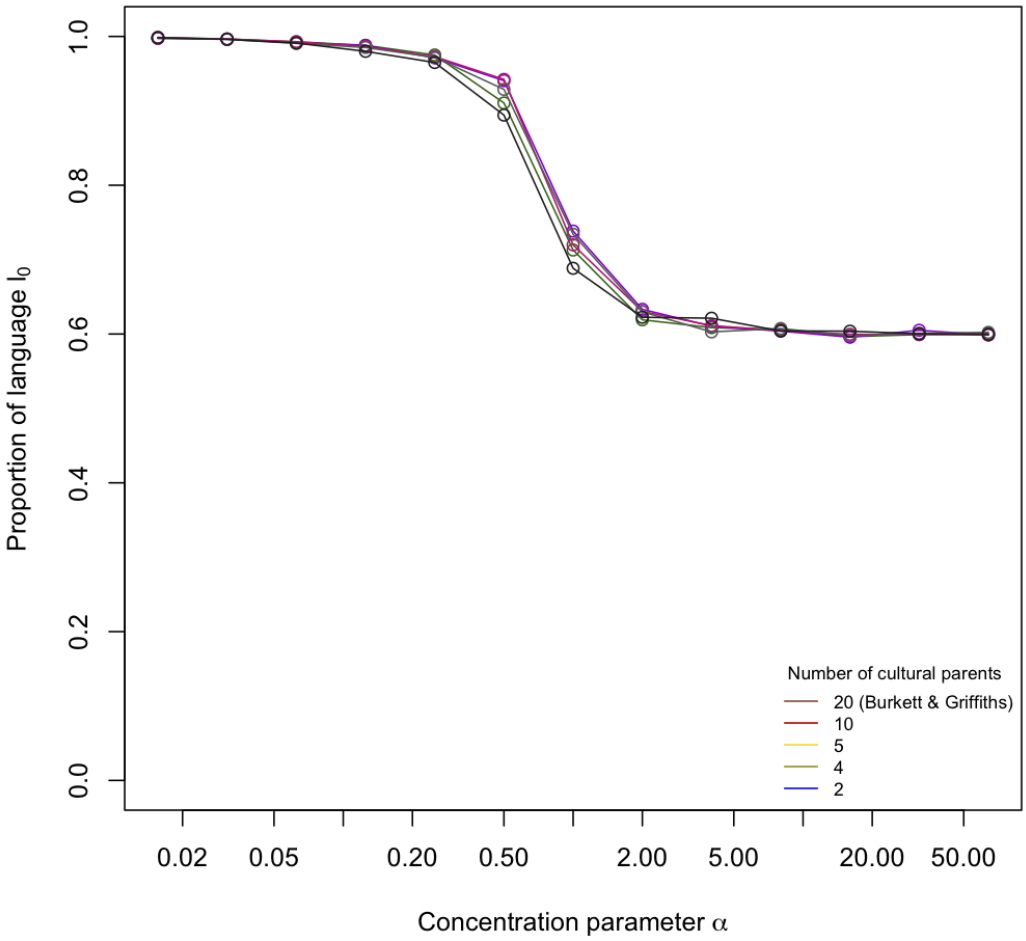

Figure 6.3: Outcome of simulations run with data being generated by 2, 4, 5, and 10 cultural parents, compared with the original results of Burkett and Griffiths (2009) where $p(w_0) = 0.6$.

Figure 6.3 shows the results of simulations run with 2, 4, 5 and 10 parents, respectively, generating a learner's input. Changing the amount of variety in the hypotheses that generate agents' input appears to have little impact on the outcome of IL. Even when $\alpha = 1$, i.e. at the point with the greatest amount of variance in the original Burkett and Griffiths (2009) model, decreasing the amount of variation does not have a consistent effect on the outcome of iterated learning, as figure 6.4 shows. For comparison, figure 6.5 shows the effect of reducing both hypothesis variety *and* the number of data items, $|d| = 5$ and 10, respectively. Again, the outcome of the simulations approximates that for $|d| = 20$, except where variance is greatest in

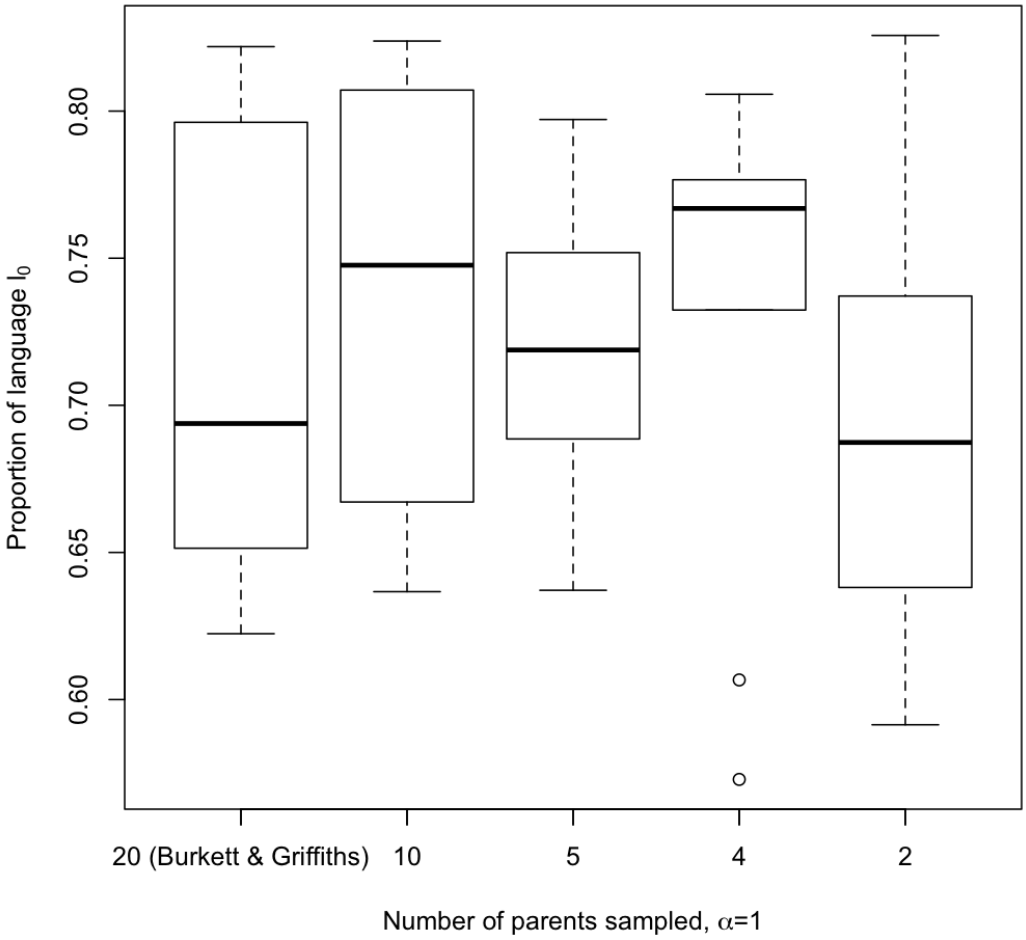

Figure 6.4: Comparison of the proportion of l_0 spoken when $\alpha = 1$ with 2, 4, 5, and 10 cultural parents. There is no clear relationship discernible between the amount of variation in hypotheses generating the data, which grows smaller as the number of cultural parents decreases, and the outcome of the iterated learning process.

the original model, i.e. for $0.25 \leq \alpha \leq 2$. Figures 6.6 and 6.7 show how the proportion of l_0 differs for different $|d|$ when α is 1 and 0.5 respectively. While input size appears to have an effect when $\alpha = 0.5$, the small sample size of ten runs make it impossible to draw any clear conclusions.

To summarize, then, neither the amount of variation in the hypotheses generating input for learners, nor the absolute amount of input learners receive, appears to have a significant impact on the outcome of the simulations. This confirms Griffiths and Kalish (2005, 2007)'s findings that Bayesian learners sampling from their posterior are not affected by a transmission bottleneck.

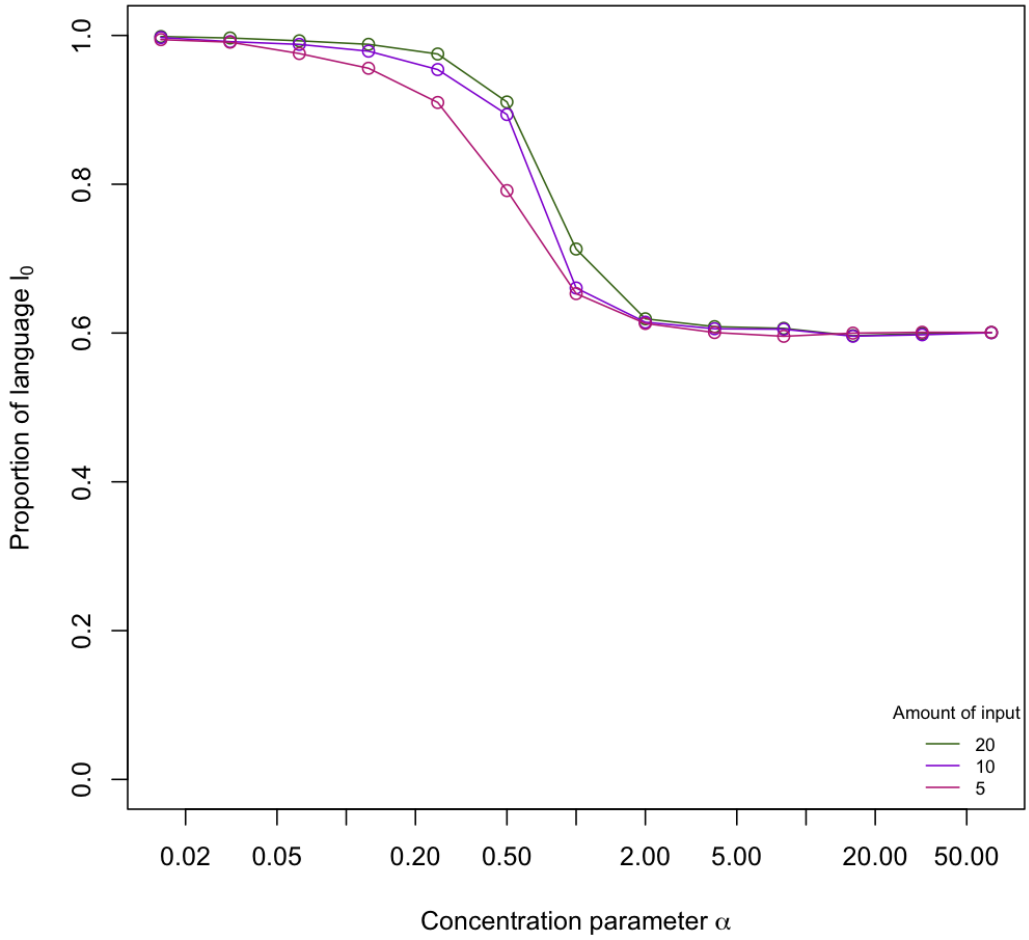

Figure 6.5: Outcome of simulations run with 5, 10 and 20 data items. the transitional interval aside, the outcome is roughly the same, regardless of the amount of input learners receive.

6.4 Modification 3: incremental Bayesian learning

In Burkett and Griffiths (2009), learners generate a hypothesis by processing all of their input at once. Anyone with children – in fact, anyone who has heard young children speak – will question whether this is an accurate approximation of how children acquire language. The CHILDES database (MacWhinney, 2000) provides numerous examples of how a child's language changes as it grows up, suggesting that, if children are engaged in some form of Bayesian learning, then this process may be incremental.[5]

To verify that Burkett and Griffiths (2009)'s results hold when agents engage in incremental Bayesian learning, the following changes were made to the model. For each learner, adults are uniformly sampled from their cohort to generate a total of $|d|$ data items. The learner then generates a hypothesis h_0 on the basis of its prior bias and the set of data containing just the

[5]Yang (2004) proposes an incremental model, with learners reassigning weights to all of the grammars within their hypothesis space after each data item they receive; his agents, however, are not Bayesian learners.

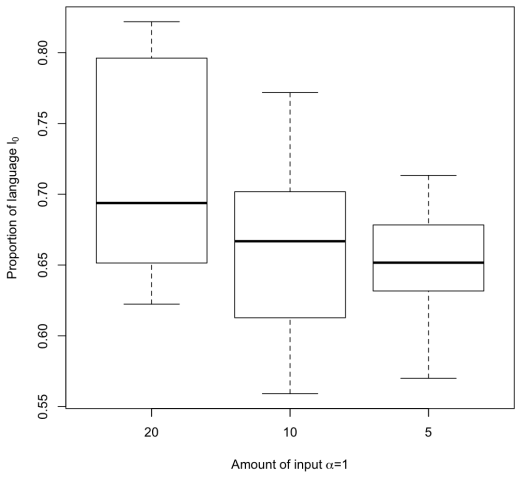

Figure 6.6: Comparison of the proportion of l_0 when $|d| =$ 20, 10, and 5; $\alpha = 1$

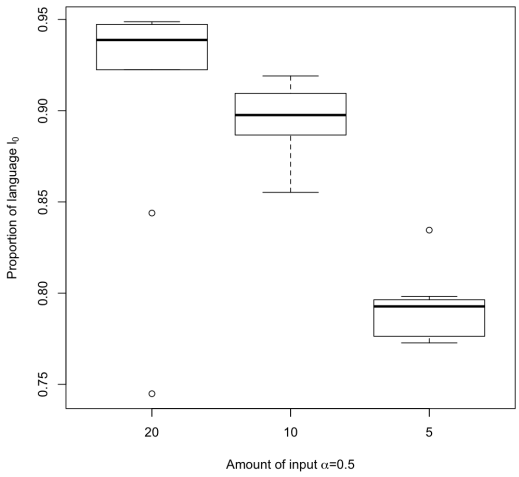

Figure 6.7: Comparison of proportion of l_0 when $|d| = 20$, 10, and 5; $\alpha = 0.5$

first item sampled, $d_0 = \{w_0\}, w_0 \in d$. For each new data item w_i, the learner generates a new hypothesis on the basis of $h_{(i-1)}$ and $d_i = \{w_i\} \cup d_{(i-1)}$:

$$p(h_i|d_i) \propto p(d_i|h_{(i-1)})p(h_{(i-1)}), \tag{6.1}$$

where the posterior $p(h_i|d_i)$ is computed using the hypothesis generated in the previous step, $p(h_{(i-1)})$, rather than the prior $p(h)$.

6.4.1 Results

Figure 6.8 presents the results of determining a hypothesis incrementally. As before, the results coincide with those of Burkett and Griffiths (2009) when α is large (i.e. $\alpha \geq 5$) or small ($\alpha \leq 0.25$). However, the transition from convergence to the prior, on the one hand, to magnification of the prior, on the other, occurs more rapidly, and for higher values of α, than in the original model. Figures 6.9 and 6.10 show how the proportion of l_0 changes during the iterated learning cycle, with values averaged over all of the agents in all ten runs of each simulation. In the case of incremental Bayesian learners, for values of α outwith the transition interval, the proportion of l_0 is virtually constant across all generations. This is somewhat difficult to explain: since the input for the first generation is generated according to the prior distribution over languages, one would expect a development similar to that of batch learners, i.e. with the first generation's posterior hypothesis reflecting the input they received. The difference may be due in part to the learners' apparent insensitivity to the amount of input they receive, shown in section 6.3: if learners can converge on a hypothesis favouring a single language with just five data items, incremental Bayesian learning might be akin to repeated sampling from data generated by the

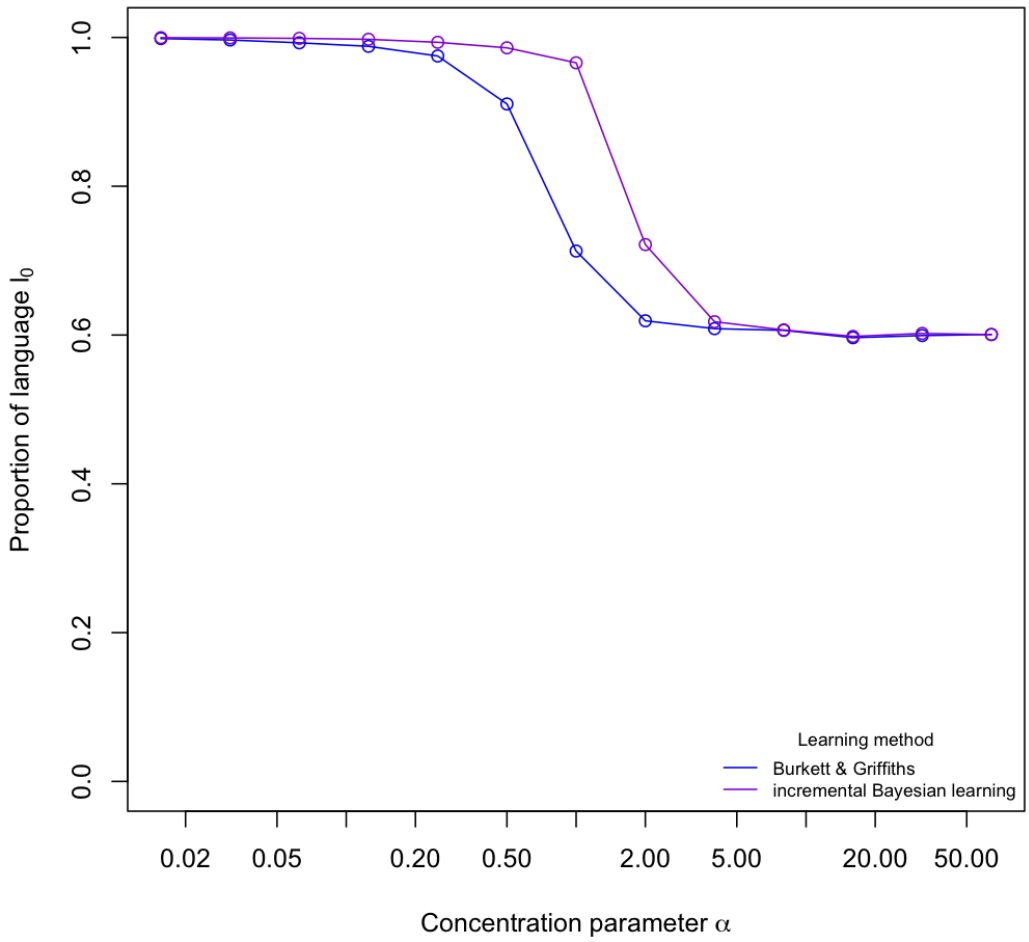

Figure 6.8: Outcome of purely incremental learning, compared to the results of Burkett and Griffiths (2009)

same hypothesis. Since a low α value suggests that the data is generated by fewer languages, learners may discard more words from the language underrepresented in their input as being erroneous.

6.5 Modification 4: horizontal transmission

As in the majority of ILMs, the data learners are provided with are generated by the previous cohort. However, it is widely assumed that children's development is influenced by their peers (e.g. Harris, 1995). This applies to children's linguistic development, too: consider the number of times adults lament that teenagers "can't talk properly", or that they don't understand what young people are saying to one another. This intuitive notion is supported by research in a number of areas. For example, according to Senghas and Coppola (2001) and Senghas et al. (2004), the increase in the complexity of Nicaraguan Sign Language – a language that has only emerged over the past three decades – can be attributed largely to the successive cohorts of new

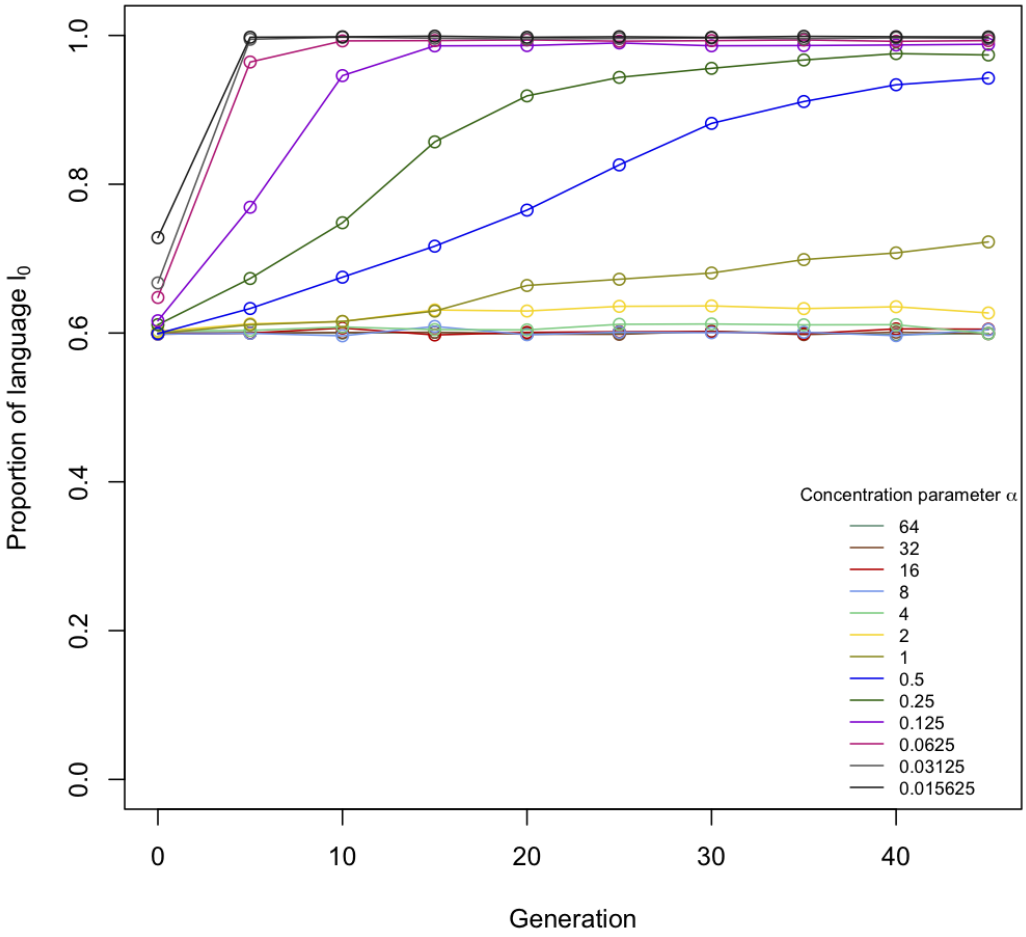

Figure 6.9: Development of the proportion of l_0, batch Bayesian learning (Burkett and Griffiths, 2009, model)

language learners adapting the language when communicating with their peers, rather than to innovation by fluent speakers.[6] Mashburn et al. (2009) report that, in hearing children, too, the language skills of their peers have a measurable impact on children's language skills. Research on bilingual education (Chesterfield et al., 1983) also suggests that the language skills of peers play an important role in shaping other learners' language skills. Vogt (2005) argues that iterated learning models that rely solely on vertical transmission cannot distinguish between regularity resulting from a transmission bottleneck imposed by the modeller and regularity that arises due to learners' creativity. He goes on to show that, in an iterated learning model, horizontal transmission can act as an implicit bottleneck.

The following simulations introduce horizontal transmission in two different ways. The first simulation uses a two-step model where learners generate two hypotheses, once after receiving data from adults and again after receiving data from other learenrs. The second model combines horizontal transmission and incremental Bayesian learning.

[6]See Senghas and Coppola (2001) for a brief history of Nicaraguan Sign Language.

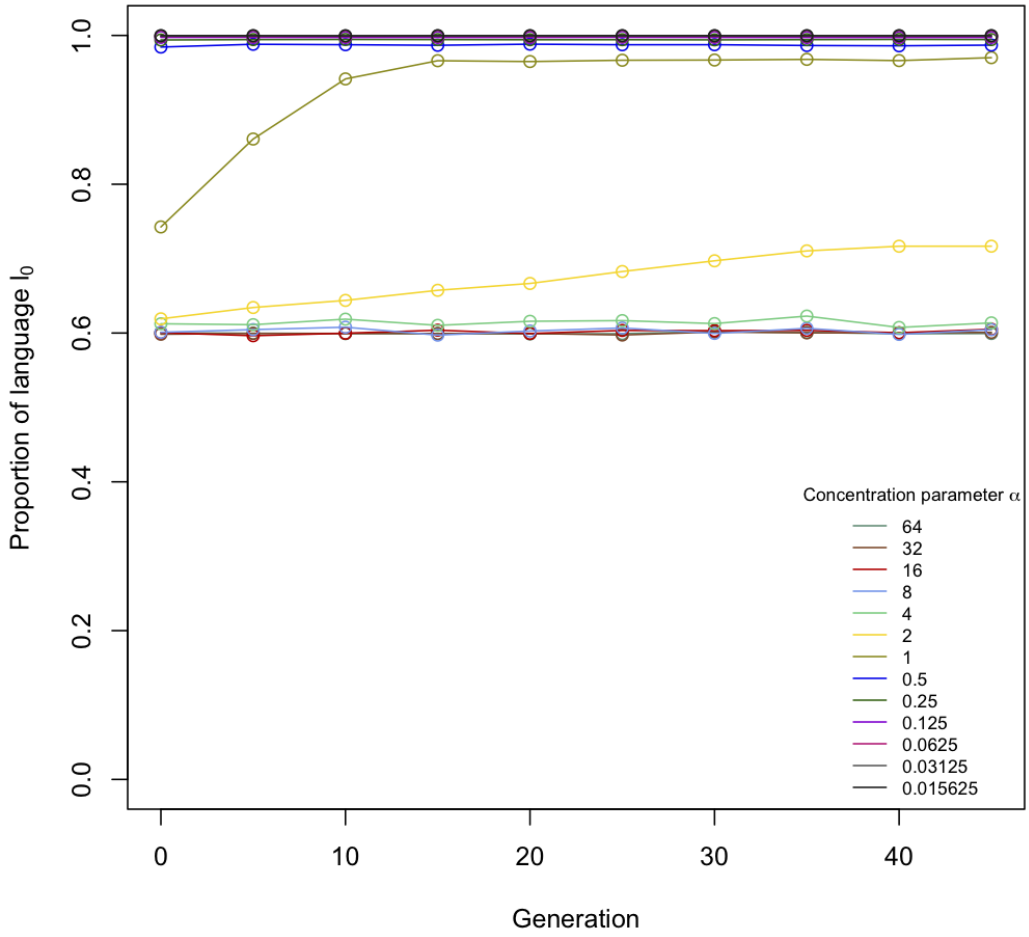

Figure 6.10: Development of the proportion of l_0, incremental Bayesian learning

6.5.1 Horizontal transmission in two steps

In this simulation, learners generate hypotheses at two points during the learning process. Initially, learners receive input from agents of the previous cohort (d_a) and generate a hypothesis concerning the distribution of languages within the population. Next, learners are called upon to generate input for each other (d_l), which is then used by learners to generate a second, final hypothesis. Two sets of simulations were run with $|d_a| = 5$ and 10, respectively, and $|d_l| = |d| - |d_a|$, with $|d| = 20$.

Results

Figure 6.11 shows the outcome of the simulations described above. Again, the overall results confirm Burkett and Griffiths (2009)'s findings: for high values of α, agents' hypotheses converge on their prior distributions over languages, whereas for small values of α, IL magnifies the starting conditions of the simulation, favouring l_0. However, for values of α around 1, the results differ,

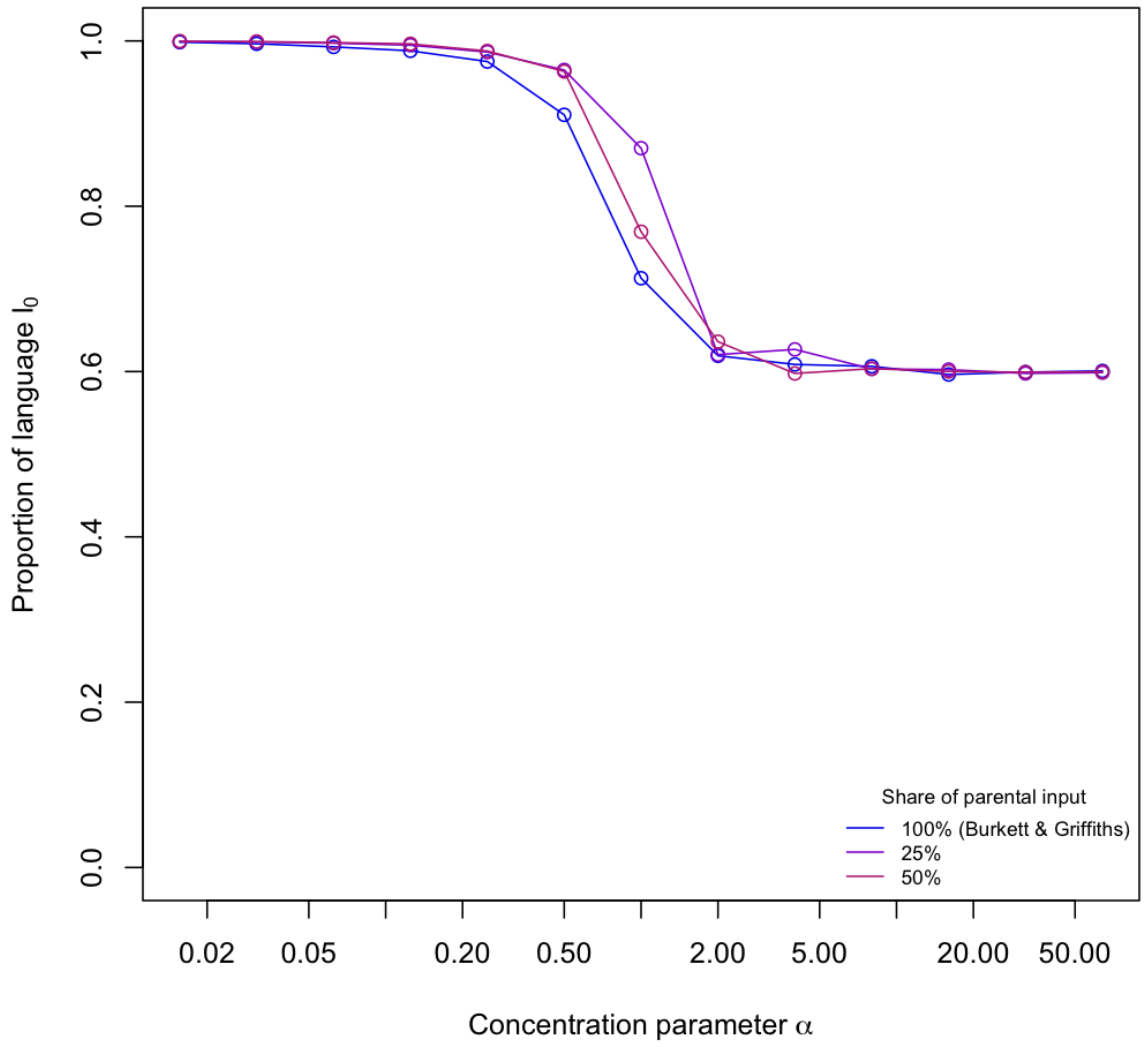

Figure 6.11: Outcome of two-step horizontal transmission, with parental input making up 100%, 50% or 25% of the input to learners

as figure 6.12 shows, albeit not greatly. Again, it is not possible to discern a general trend resulting from the amount of parental input: although figure 6.12 shows the proportion of l_0 to be higher for decreasing amounts of data from agents of the previous cohort, the proportion of l_0 is actually closer to the results of the original model when parental input makes up 25% of a learner's input than when agents from the previous cohort provide 50% of learners' input. This may, again, reflect the fact that the neither the amount of data nor the variety within the data has a significant impact on the outcome of iterated learning, as reported in section **??**.

6.5.2 Horizontal transmission with incremental Bayesian learning

In the second simulation using horizontal transmission, learners again generate a first hypothesis on the basis of initial input d_a from adult speakers. Having done so, they again only receive input

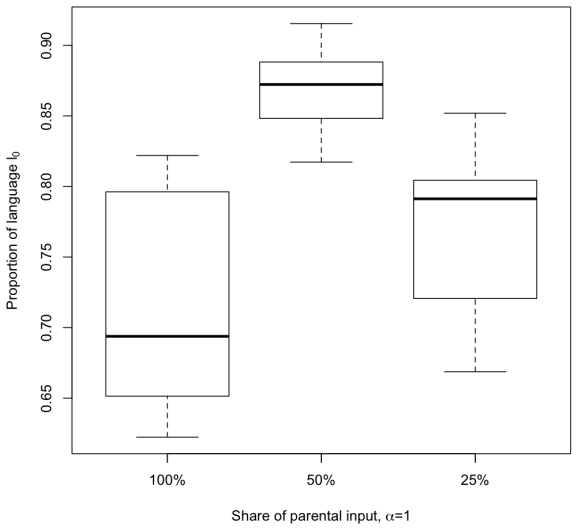

Figure 6.12: Proportion of l_0 spoken after two-step horizontal transmission with parental input making up 100%, 50% or 25% of the input to learners, $\alpha = 1$

from other learners. This time, however, they learn incrementally, generating a new hypothesis after each new data item as outlined in section 6.4.

Results

The results presented in figure 6.13 are the outcome of IL with $|d_a| = 10$. As in simulations discussed earlier, for high values of α, learners acquire a hypothesis that reflects their prior bias, whereas low α values result in hypotheses that amplify the starting conditions of the simulation, with agents learning hypotheses that prefer a single language. Again, for intermediate values of α, the results differ. As figure 6.14 shows, learning progresses in a way similar to that of pure incremental learning, with learners who have particularly high or low values for α converging to the ultimate hypothesis within a single generation. Again, this may be due to the fact that incremental learning approximates iterated learning within individual agents as a result of agents' indifference to changes in the amount of input the receive, as suggested in section 6.4.1.

6.6 Discussion

The results presented in this chapter expand on those of Burkett and Griffiths (2009): as α increases ($\alpha \to \infty$), learners' hypotheses converge to the prior G_0. For $\alpha \to 0$, on the other hand, learners generate hypotheses that amplify the initial conditions of the simulation. These results are valid in a number of scenarios that each modify one of the simplifying assumptions made by Burkett and Griffiths (2009). Together, the results confirm that the value assigned to

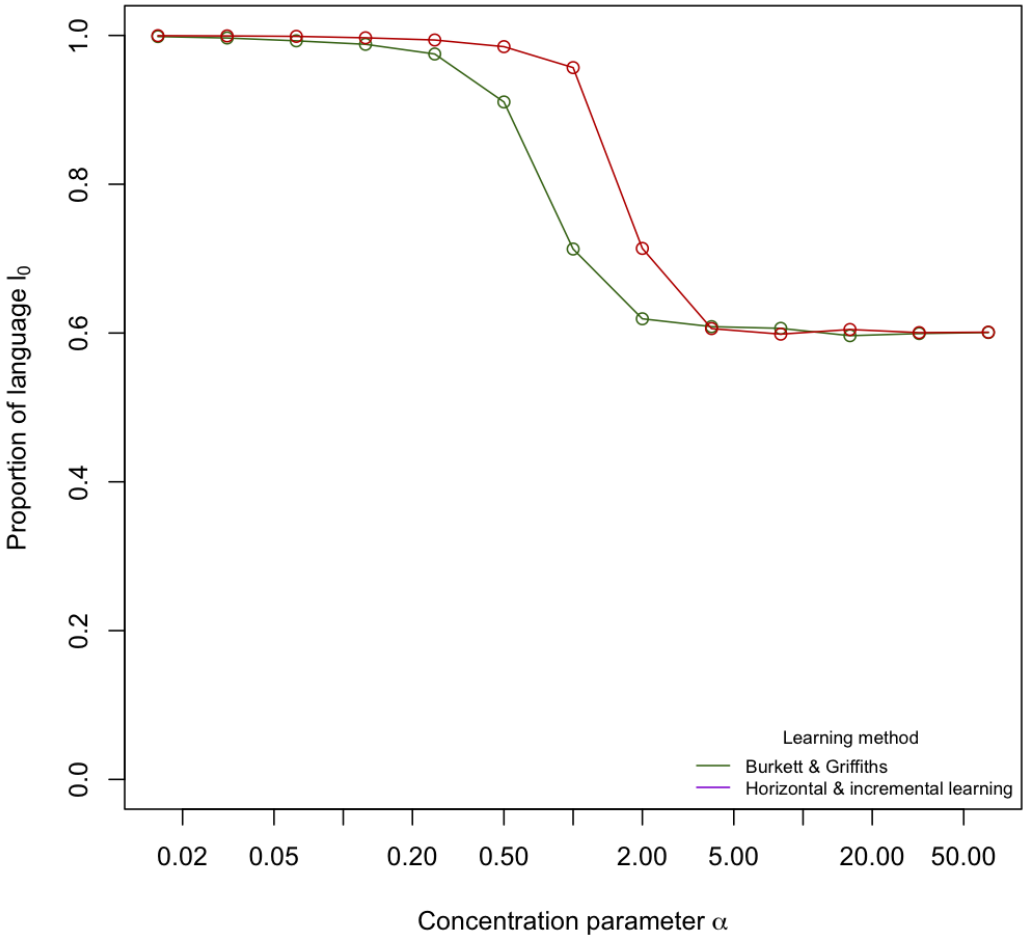

Figure 6.13: Horizontal learning with incremental Bayesian learners, $|d_a| = 10$

the concentration parameter α is crucial in determining the outcome of iterated learning with Bayesian agents under a wide variety of learning conditions. In this light, Burkett and Griffiths (2009, 2010)'s interpretation of their findings is somewhat puzzling. They appear to suggest that a high value for α is somehow a more "natural" assumption (Burkett and Griffiths, 2010, p.65):

> However, if we explicitly encode a bias in the agent towards believing that the teachers all share a single hypothesis, then we may observe results that more closely align with the initial condition.

The formulation suggests that learners ought to assume *a priori* that teachers use different languages. But why should they do that? Take the extreme case, shown in chapter 5 above, where $p(w_0) = 0$. Despite the proportion of l_0 initially only being ϵ, for $\alpha \to \infty$, learners end up using l_0 60% of the time. This is not quite the obvious default behaviour Burkett and Griffiths (2010) make it out to be. While it *may* be case that learners have high α values, the only

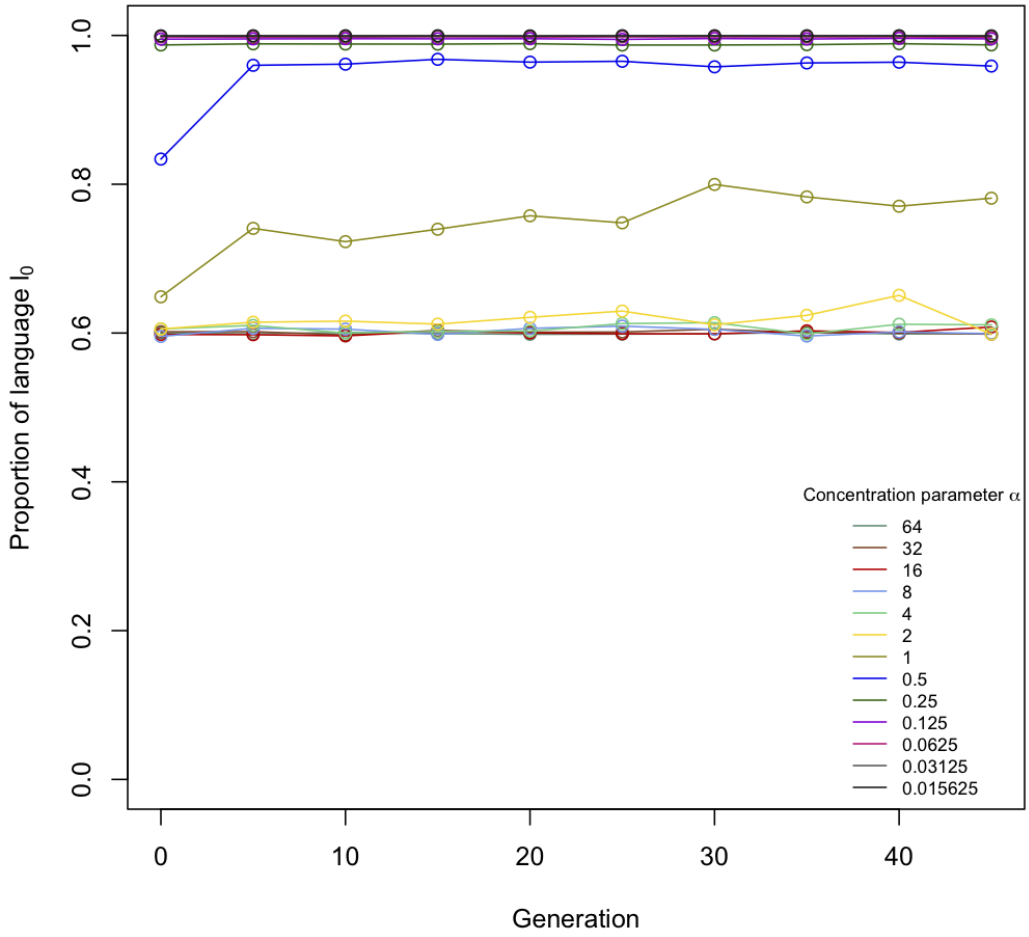

Figure 6.14: Cross-generational progress of horizontal-incremental learning for $|d_a| = 10$

argument that Burkett and Griffiths (2010) present in favour of this claim is that the outcome of simulations with high values for α more closely correlate with the outcome of learning from a single teacher. This is, I feel, a problematic line of argument. Firstly, it fails to take into consideration that under normal circumstances, humans are far more likely to have multiple teachers. To present the case of single teachers in support of assuming a particular way of learning from multiple teachers gives the impression of explaining the general case by recourse to a special one. Secondly, as Niyogi and Berwick (2009) point out, assuming transmission chains consisting of single agents per cohort restricts the dynamics of the transmission process and therefore its potential effects.

Rather than simply assuming that learners assign α a high value, it seems to me more worthwhile to ask *which value* learners assign to the concentration parameter. This is the question that will be addressed in the following chapter.

CHAPTER 7

Learning the concentration parameter α

7.1 Introduction

The models presented here so far have been simple modifications of that of Burkett and Griffiths (2009), with each model modifying one of the assumptions made in the latter. In each case, the concentration parameter α has been determined by the modeller, and determining the effect of assigning α different values has been one of the purposes of the models. If, however, one acknowledges that the value of α plays a part in determining the outcome of iterated language learning – and all of the results presented thus far strongly suggest that it does –, it seems natural to ask how the value for the concentration parameter α might be set without the intervention of a god-like modeller. Broadly speaking, there are two possible hypotheses of how the value of α might be set; I shall refer to them as the "natural" and the "cultural" origin hypothesis regarding the value of α, respectively.[1]

7.1.1 The natural origin hypothesis

The natural origins hypothesis would claim that humans have evolved a certain value for α. While this can be made to sound fairly convincing, the hypothesis runs into the same problems as hypotheses suggesting hereditary strong biases of other kinds. As discussed in chapter 2, strong domain-specific biases for particular values of α are unlikely to have evolved because the process of cultural transmission is capable of shielding weak biases from selection (Smith and

[1] It is important to note that I am putting forward both hypotheses, essentially as straw men (but with no malice aforethought); I am not aware of any work in this specific area, although Perfors and Navarro (in press) might be considered to be discussing certain aspects of the problem.

Kirby, 2008) and because the rate of language change is too fast for biological mechanisms to track sepcific linguistic features (Christiansen and Chater, 2008). It may of course be the case that α is determined by some hereditary trait within a domain other than language; what would then have to be explained, however, is how it can influence the development of (supposedly domain-specific) UG. This would lead back quite naturally to the debate between Hauser et al. (2002) and Pinker and Jackendoff (2005) concerning the nature of the language faculty.

Perfors and Navarro (in press) present an alternative natural explanation: it may be that the value for alpha is determined by the structure of the world. However, it is not entirely clear which particular aspect of the world might lead learners to assume that the concentration parameter has a particular value. Furthermore, this approach raises the question of where exactly the structure of the world is meant to lie: since it is only by perceiving that we come to know this structure, and perception is a cognitive process, this might simply amount to saying that there are ways in which different cognitive modules can affect one another.

7.1.2 The cultural origin hypothesis

If α is not determined biologically, its value must be set by some cultural means or other. For example, learners might make assumptions about the structure of their social surroundings: if everybody they interact with looks alike, they might assume that everybody speaks the same language and hence assign α a low value. This might allow a model such as that of Perfors and Navarro (in press), mentioned above, to apply to a cultural explanation of setting α as well.

An alternative might be to assume that learners are engaged in learning not only a distribution over languages but also the appropriate value for α. They would, then be faced with the task of learning a complex hypothesis (h, α) on the basis of $p(h, \alpha|d)$. One way they might go about doing so is that presented in Kemp et al. (2007), who describe how learners can acquire *overhypotheses* – essentially hypotheses that restrict the hypothesis space to be searched at a different level – using a hierarchical Bayesian model. A similar approach can be applied to learning α here. The following section describes such a model.

7.2 The model

The task learners face is to determine a complex hypothesis (h, α), consisting of a distribution over languages h and a value for α, on the basis of the distribution

$$p(h, \alpha|d) = p(d|h)p(h|\alpha)p(\alpha), \tag{7.1}$$

where $p(h|\alpha)$ is the conditional probability of h given α, $p(\alpha)$ is the prior probability of α, and $p(d|h)$ is the likelihood of d having been generated by h, defined as

$$p(d|h) = \prod_{w \in d} (\sum_{l} p(l|h)p(w|l)). \tag{7.2}$$

Here, $p(w|l)$ defined as in equation 5.5. As in the Burkett and Griffiths (2009) model, determining $p(l|h)$ exactly is not possible but must instead be sampled as part of the process described below.

Each agent samples a complex hypothesis (h, α) from its posterior distribution $p(h, \alpha|d)$ using the Metropolis-Hastings algorithm based on the Chinese Restaurant Process: initially, the agent is provided with $|d|$ data items. It then determines a value for α by sampling a value for $\ln(\alpha)$ from a Gaussian distribution with mean $\mu = 1$ and standard deviation $\sigma = 1$. This method has the advantage of providing a probability density function for α in the shape of a log-normal distribution, which can be used to determine $p(\alpha)$. Next, the agent generates an assignment of words w to clusters c according to the distribution

$$p(c|\alpha, \mathcal{C}) = \begin{cases} \frac{n_c}{N+\alpha} & c \text{ is an existing cluster} \\ \frac{\alpha}{N+\alpha} & c \text{ is a new cluster} \end{cases}, \tag{7.3}$$

where $\mathcal{C}$ is the set of all clusters, n_c is the number of items assigned to cluster c and N is the number of items assigned so far. For K clusters with data items assigned, $N = \sum_{1 \leq k \leq K} n_k$. Items are assigned to clusters until $N = |d|$. The agent then assigns a language l to the cluster according to the distribution

$$p(l) \propto G_0(l) \tag{7.4}$$

Note that both the assignments of "data items" and of languages to clusters are independent of $w \in d$, i.e. the data the agent has actually been provided with; they are essentially being generated randomly. Together, the sampled values for α and assignments allow the agent to generate a hypothesis according to equation 5.9. Next, the agent determines $p(d|h)$ according to equation 7.2. With this done, all of the elements of equation 7.1 are in place and the agent can calculate $p(h, \alpha|d)$. This entire process (with the exception of being provided with data) is repeated to determine a second complex hypothesis defined by $p(h', \alpha'|d)$. The agent adopts the complex hypothesis (h', α') if $p(h', \alpha'|d) > p(h, \alpha|d)$; if this condition is not met, it adopts (h', α') with probability $\frac{\alpha'}{\alpha}$.

The process of generating a new complex hypothesis is repeated 1000 times per agent, with the outcome of the final iteration being adopted by the agent as its hypothesis.

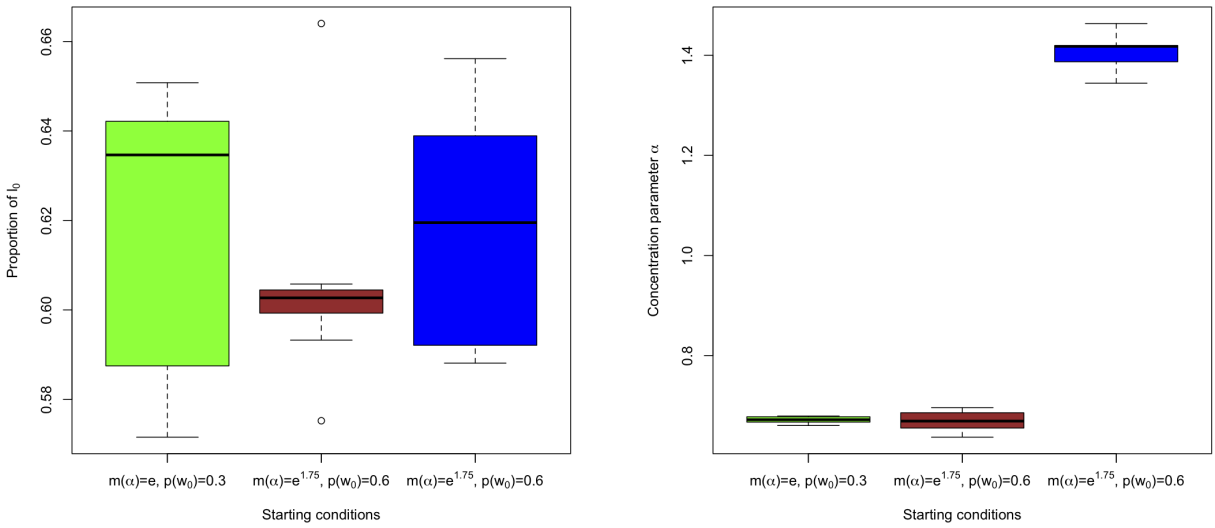

Figure 7.1: Value of α and proportion of l_0 spoken after learning a hypothesis for (h, α) under different conditions; results from simulations with the same initial conditions are coloured the same. $m(\alpha)$ refers to the mean of the prior (log-normal) distribution for α.

7.3 Results

Figure 7.1 shows the outcome of learning a complex hypothesis (h, α). As one can see on the left, learners appear to converge to their prior distribution over languages, just as they did for large values of α previously. However, as one can see on the right, the values of α are considerably lower than the Burkett and Griffiths (2009) model (and the variations thereof presented above) would lead us to expect given the proportion of l_0 being spoken. The proportion of l_0 in the initial data, $p(w_0)$, does not appear to have any significant impact on the outcome of iterated learning. Figure 7.1 suggests that it may have influenced the amount of variance in the proportion of l_0 for $p(w_0) = 0.3$, but the small number of runs being compared and the amount of variance in the outcomes for individual agents make it difficult to draw any hard and fast conclusions.

Although values for α do not converge to the prior, the agents' prior bias regarding α is reflected in the mean value of α, as figure 7.1 shows: the mean value for α in the condition where $m(\alpha) = e^{1.75}$ is roughly 2.1 times that of the condition where $m(\alpha) = e$, which is approximately the ratio of the priors' means, $\frac{e^{1.75}}{e}$. This result might indicate the onset of convergence to a value for α smaller than the prior, although the large variance in the results mean that such a claim is problematic (or possibly premature). Nevertheless, this is potentially an interesting result: as mentioned in section 6.6, Burkett and Griffiths (2009) seem to assume that learners have a default preference for large values of α. If the tendency for α to assume values smaller than $m(\alpha)$ were borne out in future simulations, then that assumption may have to be revised.

Surprisingly, iterated learning does not seem to have an effect on the learning outcome,

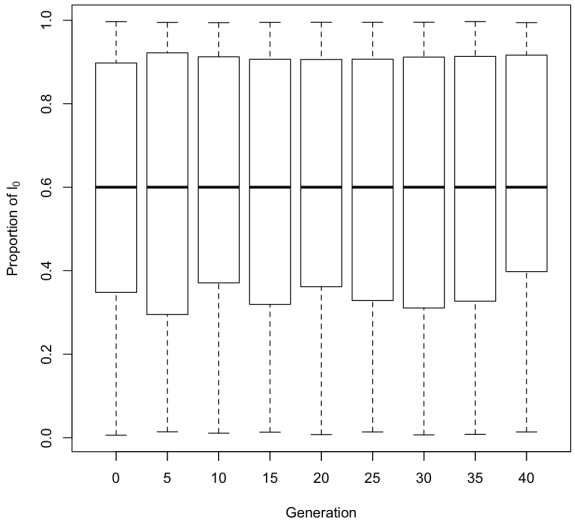

Figure 7.2: Development of $p(l_0)$ over time, $m(\alpha) = e, p(w_0) = 0.3$

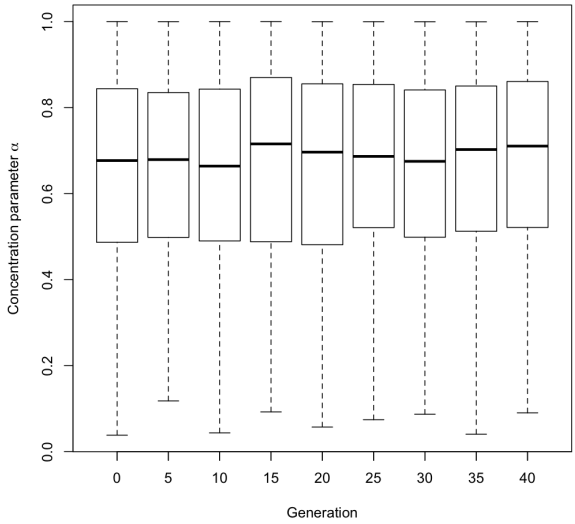

Figure 7.3: Development of α over time, $m(\alpha) = e, p(w_0) = 0.3$

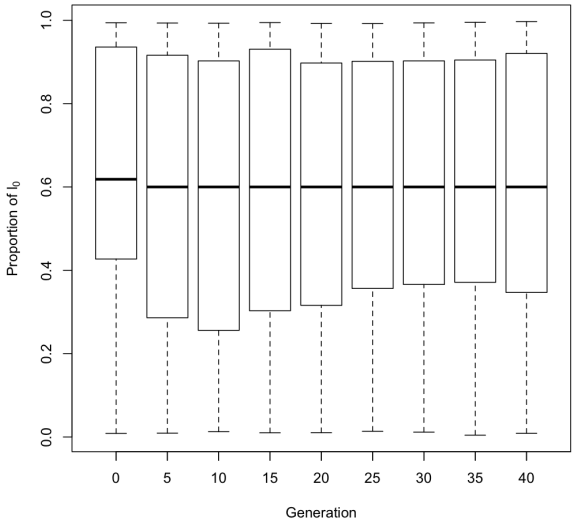

Figure 7.4: Development of $p(l_0)$ over time, $m(\alpha) = e, p(w_0) = 0.6$

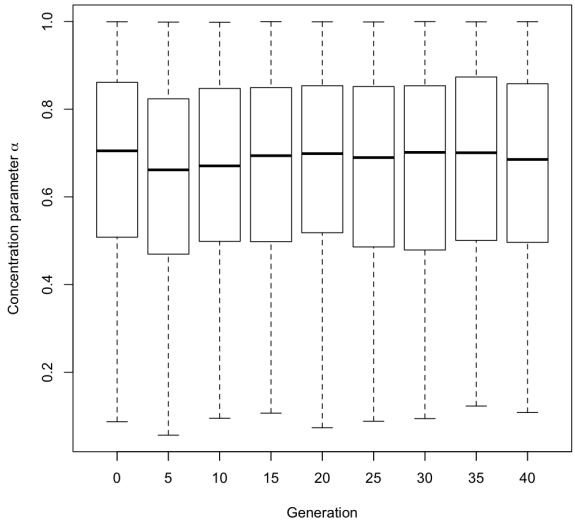

Figure 7.5: Development of α over time, $m(\alpha) = e, p(w_0) = 0.6$

either: as figures 7.2 to 7.7 show, throughout the learning process, both α and the share of l_0 vary greatly, with no apparent convergence. Thus, although agents appear to be converging to the prior if one looks merely at the mean share of l_0, in fact agents seem to be opting for an extremely wide variety of hypotheses that often reflect neither their prior biases nor the starting conditions of the simulation. The scatterplots in figures 7.8 to 7.13 confirm that there is no correlation between the value for α and the hypothesis regarding the proportion of l_0 within the population. It is clearly visible that agents tend to have a preference for a hypothesis reflecting their prior bias regarding the distribution of languages, but this does not change over time: the complex hypotheses agents acquire after 45 cohorts are just as widely distributed as those seen

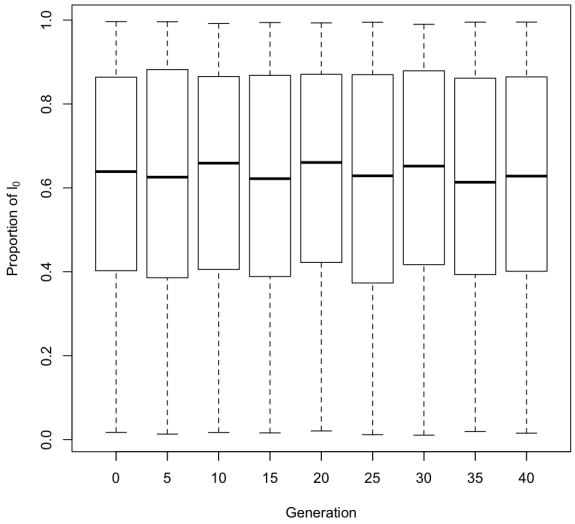

Figure 7.6: Development of $p(l_0)$ over time, $m(\alpha) = e^{1.75}, p(w_0) = 0.6$

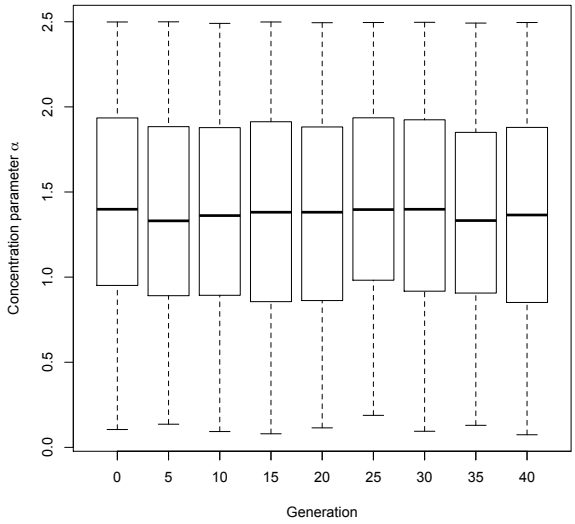

Figure 7.7: Development of α over time, $m(\alpha) = e^{1.75}, p(w_0) = 0.6$

after a single cohort.

7.4 Discussion

The results presented here are rather surprising, since Kemp et al. (2007) show that hierarchical Bayesian models of the kind presented here are capable of making the kind of inferences that would be necessary to learn (h, α), and Perfors et al. (2006, 2010) show how a similar model is capable of selecting a hypothesis made up of a type of grammar and a specific instantiation of such a grammar.[2] A possible explanation for this discrepancy, especially between Kemp et al. (2007)'s results and those presented here, is the amount of learning each agent undertook: whereas agents in this model select a hypothesis after 1000 iterations of the sampling process, Kemp et al. (2007)'s agents ran 50000 iterations. It may be the case that the agents in the current model would also have shown greater signs of convergence after that number of iterations; unfortunately, computational limitations made it impossible to run that many iterations.[3]

One result – or possibly a hint of a result – is that agents may be converging on α values below $m(\alpha)$. This incipient convergence is influenced by the learners' prior bias regarding α, rather than being an absolute preference for a particular value. It may be that this is a spurious effect which would disappear after a few thousand more iterations of the sampling process; it might, however, be the beginning of learners converging on low values of α. If this could be shown

[2] It should be noted that both models are concerned with a single generation of learners, not iterated learning.

[3] A single run of the model described in section 7.2 took approximately 55 minutes to complete 1000 iterations on the hardware available. Ten runs of the model with a single set of initial conditions and 50000 iterations would thus have taken roughly 19 days to complete.

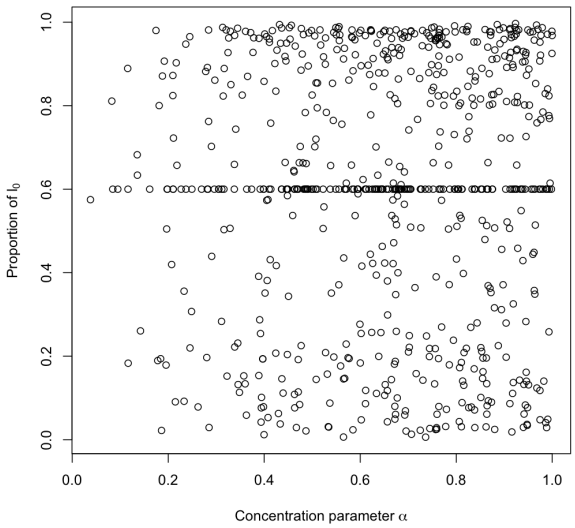

Figure 7.8: Composition of (h, α) in generation 1, $m(\alpha) = e, p(w_0) = 0.3$

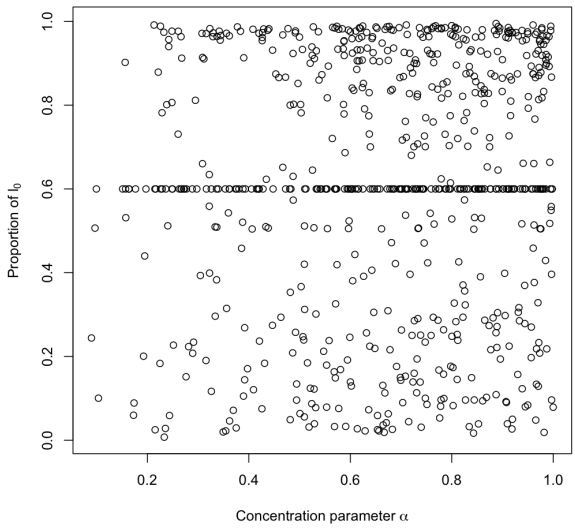

Figure 7.9: Composition of (h, α) in generation 46, $m(\alpha) = e, p(w_0) = 0.3$

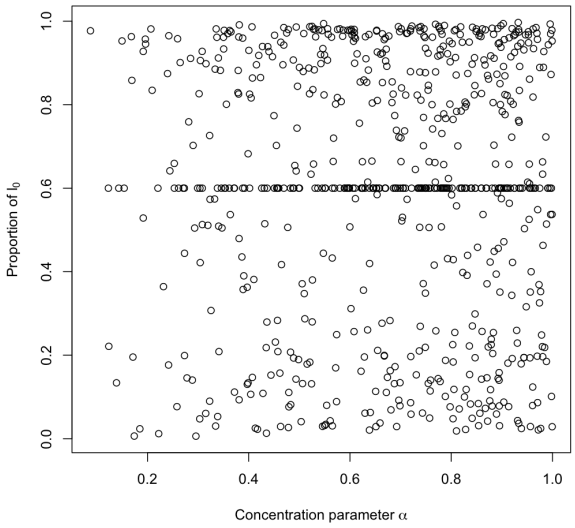

Figure 7.10: Composition of (h, α) in generation 1, $m(\alpha) = e, p(w_0) = 0.6$

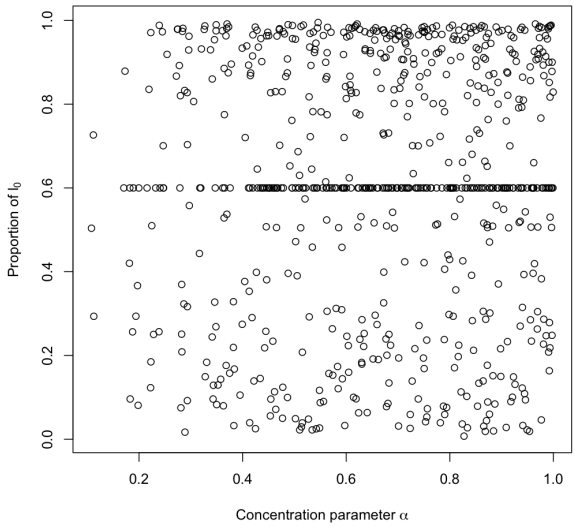

Figure 7.11: Composition of (h, α) in generation 46, $m(\alpha) = e, p(w_0) = 0.6$

to be the case, e.g. by running simulations with more iterations of the sampling process, this would mean that Burkett and Griffiths (2009, 2010)'s assumption regarding learners' preference for high values of α may need revising. The latter's results would still hold, of course; however, depending on the value for α that learners converged on, the fact that they converge on the prior distribution over languages for $\alpha \to \infty$ may be less significant for modelling the human language acquisition process than previously assumed.

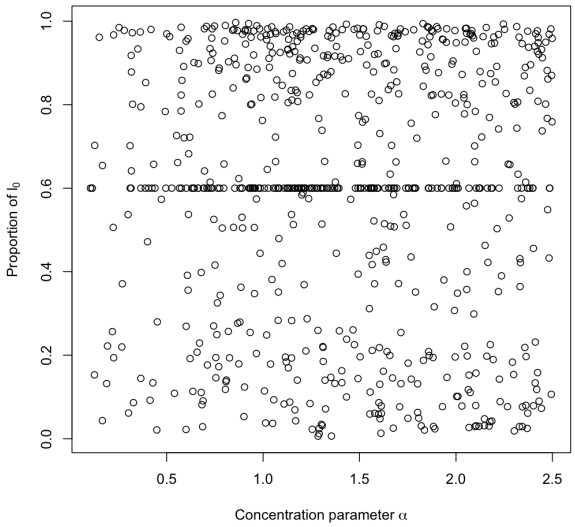

Figure 7.12: Composition of (h, α) in generation 1, $m(\alpha) = e, p(w_0) = 0.3$

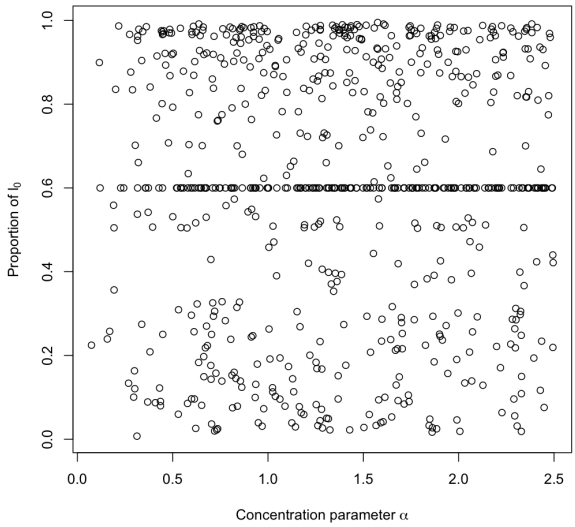

Figure 7.13: Composition of (h, α) in generation 46, $m(\alpha) = e^{1.75}, p(w_0) = 0.6$

CHAPTER 8

Discussion and conclusion

8.1 Discussion

The results presented in the previous two chapters fall into two broad categories. Chapter 6 presents a series of results showing that the general findings of Burkett and Griffiths (2009, 2010) – i.e. that small values of α allow the initial input learners receive to affect their hypotheses, leading to hypotheses that magnify imbalances within the input, whereas large α values result in agents' hypotheses converging to their prior bias regarding the distribution of languages within the population – apply to a broader range of scenarios than previously known. Chapter 7, on the other hand, is an attempt to go beyond the original model and determine which α value learners select if it is not provided by the modeller but must instead be learned along with the distribution over languages in the population.

Section 6.3 shows that Burkett and Griffiths (2009)'s results are insensitive to the amount of variety within the hypotheses that generate learners' input and the absolute amount of data they receive. It makes little difference whether data is generated by 2, 4, 5, 10, or 20 adults from the previous cohort, or whether they receive 5, 10, or 20 data items as input: the outcome of IL closely resembles that of the original model. In section 6.4, learners engage in incremental Bayesian learning. Again, Burkett and Griffiths (2009)'s results are largely confirmed. However, for low values of α, learners converge to a hypothesis favouring a single language far more quickly than in the original model. As stated in section 6.4.1, this may be due to agents' indifference to the amount of input they receive. Allowing horizontal as well as vertical transmission as in section 6.5 does not alter the outcome of IL greatly either, although horizontal-incremental

learning exhibits the same accelerated convergence to hypotheses favouring a single language as pure incremental learning. One scenario, that of populations consisting of agents with different α values (section 6.2), may at first seem to challenge Burkett and Griffiths (2009)'s results; however, as discussed in section 6.2.1, a possible explanation for this is that agents in this scenario are not behaving like rational Bayesian agents since they are essentially selecting a hypothesis on the assumption that the data they receive as input was generated by a homogeneous population.

Although their *findings* are shown to be more widely applicable than was previously known to be the case, it does not follow that Burkett and Griffiths (2009, 2010)'s *interpretation* of their results also gain additional support. In fact, their interpretation is somewhat curious. They appear to be suggesting that, by default, learners have a high α value because when that is so, the results of learning from multiple teachers coincide with those for single-agent transmission chains, presented by Griffiths and Kalish (2005, 2007). At the same time they acknowledge that learning from a single teacher is essentially a special case of learning from multiple teachers. This creates the impression that the special case is being used to support a particular interpretation of the general case. Instead, it seems more reasonable to say that whether the findings of Griffiths and Kalish (2005, 2007) apply to human language learning depends on whether learners can be assumed to have a high α value. In addition, the results of chapter 6 show that there is an interval for values of α of approximately $0.25 \leq \alpha \leq 2$ where the outcome of IL is less clearcut. It is in this interval that the modified scenarios also show the greatest degree of divergence from the original model. Yet Burkett and Griffiths (2009, 2010) barely discuss this facts, limiting themselves instead to the effects of $\alpha \to \infty$ and $\alpha \to 0$.

The model presented in chapter 7 attempts to show how learners might acquire a particular value for α, rather than simply assuming it is given, as Burkett and Griffiths (2009, 2010) do. Learners acquire a complex hypothesis (h, α), consisting of both a hypothesis about the distribution of languages within the population and a value for α. The results of the simulations run with this model are far less conclusive than those of chapter 6. There appears to be no convergence whatsoever, either on a particular value for α or a particular distribution over languages h. One possible explanation for this is that learners have not had enough time to "mull their data over", that is, the number of iterations they performed to sample a complex hypothesis (h, α) was not large enough. However, increasing the number of iterations to that of similar simulations, e.g. Kemp et al. (2007), was not possible for practical reasons. Thus it may be the case that allowing agents 50000 iterations, as opposed to a "paltry" 1000, might have resulted in convergence to a particular complex hypothesis; the results presented are sadly not sufficient to allow such a conclusion to be drawn.

One particular aspect of the results from chapter 7 is worth noting, however. It was shown that although there was a great deal of variance, the mean value for α within the population was significantly below the mean of the prior for α, yet there appeared to be a correlation between the mean $m(\alpha)$ of the prior and the mean of α within the population. This suggests that the value of α is sensitive to iterated learning and that learners may indeed be able to acquire a complex hypothesis of both α and a distribution over languages. It might also suggest that learners do not, as Burkett and Griffiths (2009, 2010) assume, have a "natural" preference for high values of α. Whether that is indeed the case, however, depends on whether this apparent effect is maintained in a simulation that produces more conclusive results, e.g. in a simulation with, say, 10000 or 50000 iterations of the sampling process. For such a result to then be applicable to language learning, will also depend on whether the log-normal prior chosen for α reflects the kind of prior biases human learners have.

8.2 Conclusion

This dissertation has provided evidence that Bayesian learners sampling from their posterior distribution converge to a hypothesis reflecting their prior bias under a wide variety of conditions. It has also shown how these results might be expanded to determine whether the conditions necessary for convergence to the prior to occur can emerge as the result of iterated learning. Although the results with regards to the latter are inconclusive, they do at least hint that iterated learning may indeed influence the variables that determine whether such convergence takes place or not, thereby allowing Bayesian learners to acquire the necessary complex hypotheses. However, more research is needed to determine whether this possibility is real or only a temporary, and basically coincidental, state of the sampling algorithm used here. Simulations with a greater number of iterations of the sampling process would, hopefully, provide more conclusive answers.

Bibliography

Anderson, S. R. and Lightfoot, D. (2002). *The Language Organ*. Cambridge University Press, Cambridge.

Barr, D. (2004). Establishing conventional communication systems: Is common knowledge necessary? *Cognitive Science*, 28(6):937–962.

Bartlett, F. C. (1932). *Remembering*. Macmillan, Oxford.

Beppu, A. and Griffiths, T. (2009). Iterated learning and the cultural ratchet. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2089–2094, Austin, TX. Cognitive Science Society.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54.

Burkett, D. and Griffiths, T. L. (2009). Exploring Multilingual Hypotheses in Iterative Learning. Technical report, University of California, Berkeley.

Burkett, D. and Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In Smith, A. D. M., Schouwstra, M., de Boer, B., and Smith, K., editors, *The Evolution of Language: Proceedings of the 8th International Conference (EVOLANG8)*, pages 58–65. World Scientific Publishing Company.

Chater, N. and Christiansen, M. H. (2010). Language Acquisition Meets Language Evolution. *Cognitive Science*, 34:1131–1157.

Chater, N. and Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–44.

Chesterfield, R., Hayes-Latimer, K., Chesterfield, K. B., Chávez, R., and Chavez, R. (1983). The Influence of Teachers and Peers on Second Language Acquisition in Bilingual Preschool Programs. *TESOL Quarterly*, 17(3):401.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press, Cambridge (MA).

Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *The Behavioral and brain sciences*, 31(5):489–508; discussion 509–58.

Crystal, D. (2000). *Language Death*. Cambridge University Press, Cambridge.

Dabrowska, E. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native English speakers. *Language Sciences*, 28(6):604–615.

Dediu, D. (2008). Causal Correlations Between Genes and Linguistic Features – the Mechanism of Gradual Language Evolution. In Smith, A. D. M., Smith, K., and Ferrer i Cancho, R., editors, *The Evolution of Language - Proceedings of the 7th International Conference (EVOLANG7)*, pages 83–90, Singapore. World Scientific Publishing Co. Pte. Ltd.

Dediu, D. (2009). Genetic biasing through cultural transmission: do simple Bayesian models of language evolution generalize? *Journal of Theoretical Biology*, 259(3):552–61.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–48; discussion 448–494.

Ferdinand, V. and Zuidema, W. (2009). Thomas' theorem meets Bayes' rule : a model of the iterated learning of language. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 1786–1791.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14(3):119–30.

Fitch, W. T., Hauser, M. D., and Chomsky, N. (2005). The evolution of the language faculty: clarifications and implications. *Cognition*, 97(2):179–210; discussion 211–25.

Frank, M. C., Goodman, N. D., and Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–85.

Frigyik, B. A., Kapila, A., and Gupta, M. R. (2010). Introduction to the Dirichlet Distribution and Related Processes. Technical report, Department of Electrical Engineering, University of Washington, Seattle WA.

Gerken, L., Wilson, R., and Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, 32(2):249–268.

Godfrey-Smith, P. (2007). Innateness and Genetic Information. In Carruthers, P., Laurence, S., and Stich, S., editors, *The Innate Mind: Foundations and the Future*. Oxford University Press, Oxford.

Gökaydin, D., Ma-Wyatt, A., Navarro, D., and Perfors, A. (in press). Humans use different statistics for sequence analysis depending on the task. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Gold, M. E. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7(5):219–224.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5):431–6.

Goodman, N. (1967). The epistemological argument. *Synthese*, 17(1):23–28.

Griffiths, T. L. (2011). Rethinking language: how probabilities shape the words we use. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10):3825–6.

Griffiths, T. L. and Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 827–832. Citeseer.

Griffiths, T. L. and Kalish, M. L. (2007). Language Evolution by Iterated Learning With Bayesian Agents. *Cognitive Science*, 31:441–480.

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review*, 102(3):458–489.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, 298(5598):1569–79.

Hawking, S. W. (1988). *A Brief History of Time*. Bantam, London.

Hoff, E. (2003). The specificity of environmental influence: socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, 74(5):1368–78.

Hurford, J. R. (2000). Social transmission favours linguistic generalization. In Knight, C., Hurford, J. R., and Studdert-Kennedy, M., editors, *The evolutionary emergence of language: Social function and the origins of linguistic form*, pages 324–352. Cambridge University Press, Cambridge.

Jackendoff, R. (2003). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

Jackendoff, R. and Pinker, S. (2005). The nature of the language faculty and its implications for evolution of language (Reply to Fitch, Hauser, and Chomsky). *Cognition*, 97(2):211–225.

Kalish, M., Griffiths, T., and Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294.

Kemp, C., Perfors, A., and Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, 10(3):307–21.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8(2):185–215.

Kirby, S. (2007). The evolution of meaning-space structure through iterated learning. In Lyon, C., Nehaniv, C., and Cangelosi, A., editors, *Emergence of communication and language*, pages 253–267. Springer, Berlin.

Kirby, S., Cornish, H., and Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10681–6.

Kirby, S., Dowman, M., and Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12):5241–5245.

Kirby, S. and Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer, London.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago University Press, Chicago.

Ladd, D. R., Dediu, D., and Kinsella, A. R. (2008). Languages and Genes: Reflections on Biolinguistics and the Nature – Nurture Question. *Biolinguistics*, 2(1):114–126.

Ladefoged, P. (1984). 'Out of chaos comes order': Physical, biological and structural patterns in phonetics. In *Proceedings of the Tenth International Congress of Phonetic Science*, pages 83–95.

Lewis, M. P., editor (2009). *Ethnologue: Languages of the World*. SIL International, Dallas (TX), 16th edition.

MacWhinney, B. (2000). *The CHILDES project. Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah (NJ), 3rd edition.

Mameli, M. and Bateson, P. (2006). Innateness and the sciences. *Biology and Philosophy*, 21(2):155–188.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco (CA).

Mashburn, A. J., Justice, L. M., Downer, J. T., and Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development*, 80(3):686–702.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–11.

Mesoudi, A. and Whiten, A. (2008). Review. The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1509):3489–501.

Mesoudi, A., Whiten, A., and Dunbar, R. (2006). A bias for social information in human cultural transmission. *British Journal of Psychology*, 97(Pt 3):405–23.

Monaghan, P., Chater, N., and Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96(2):143–82.

Navarro, D. and Perfors, A. (2010). The Chinese restaurant process. Technical report, University of Adelaide.

Niyogi, P. and Berwick, R. C. (2009). The proper treatment of language acquisition and change in a population setting. *Proceedings of the National Academy of Sciences of the United States of America*, 106(25):10124–9.

Nowak, M. a., Plotkin, J. B., and Jansen, V. a. (2000). The evolution of syntactic communication. *Nature*, 404(6777):495–8.

Pancsofar, N. and Vernon-Feagans, L. (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology*, 27:571 – 587.

Perfors, A. and Navarro, D. (in press). Language evolution is shaped by the structure of the world: An iterated learning analysis. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Perfors, A., Tenenbaum, J., and Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 275–281, Austin, TX. Cognitive Science Society.

Perfors, A., Tenenbaum, J. B., Griffiths, T. L., and Xu, F. (2011). A tutorial introduction to Bayesian models of cognitive development. *Cognition*, 120(3):302–21.

Perfors, A., Tenenbaum, J. B., and Regier, T. (2010). The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Pinker, S. and Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13:707–784.

Pinker, S. and Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, 95(2):201–36.

Pullum, G. K. and Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):9–50.

Putnam, H. (1967). The 'Innateness Hypothesis' and Explanatory Models in Linguistics. *Synthese*, 17:12–22.

Quine, W. V. O. (1975). *Word and Object*. MIT Press, Cambridge (MA).

Rowland, C. F., Pine, J. M., Lieven, E. V. M., and Theakston, A. L. (2003). Determinants of acquisition order in wh-questions: re-evaluating the role of caregiver speech. *Journal of Child Language*, 30:609–635.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–8.

Sampson, G. (2005). *The 'Language Instinct' Debate*. Continuum, London, New York, 2nd edition.

Sanborn, A. N. and Griffiths, T. L. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Scholz, B. C. and Pullum, G. K. (2002). Searching for arguments to support linguistic nativism. *The Linguistic Review*, 19:185–223.

Senghas, a. and Coppola, M. (2001). Children creating language: how Nicaraguan sign language acquired a spatial grammar. *Psychological Science*, 12(4):323–8.

Senghas, A., Kita, S., and Ozyürek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. *Science*, 305(5691):1779–82.

Shultz, T. R. (2007). The Bayesian revolution approaches psychological development. *Developmental Science*, 10(3):357–364.

Smith, A. D. M. (2001). Establishing communication systems without explicit meaning transmission. In Kelemen, J. and Sosík, P., editors, *ECAL01*, pages 381–390, Prague. Springer.

Smith, A. D. M. (2005). The Inferential Transmission of Language. *Adaptive Behavior*, 13(4):311–324.

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Science*, 14(1):65–84.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 697–702.

Smith, K. and Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1509):3591–603.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–86.

Sperber, D. (2000). Metarepresentations in an evolutionary perspective. In Sperber, D., editor, *Metarepresentations: A Multidisciplinary Perspective*, pages 117–137. Oxford University Press, New York.

Sperber, D. and Origgi, G. (2010). A pragmatic perspective on the evolution of language. In Larson, R. K., Déprez, V., and Yamakido, H., editors, *The Evolution of Human Language: Biolinguistic Perspectives*, pages 124–131. Cambridge University Press, Cambridge.

Steels, L. (2000). The puzzle of language evolution. *Kognitionswissenschaft*, 8(4):143–150.

Swarup, S. and Gasser, L. (2009). The Iterated Classification Game: A New Model of the Cultural Transmission of Language. *Adaptive Behavior*, 17(3):213–235.

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., and Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 332(6033):1054–9.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, 331(6022):1279–85.

Theakston, A. L., Lieven, E. V. M., and Tomasello, M. (2003). The role of the input in the acquisition of third person singular verbs in English. *Journal of Speech, Language, and Hearing Research*, 46(4):863–77.

Thomas, M. (2002). Development of the concept of "the poverty of the stimulus". *The Linguistic Review*, 19(1-2):51–71.

Tomasello, M. (2003). *Constructing a Language: a Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge (MA).

Tomasello, M. (2005). Beyond formalities: The case of language acquisition. *The Linguistic Review*, 22:183–197.

Ventureyra, V., Pallier, C., and Hoo, H.-Y. (2004). The loss of first language phonetic perception in adopted Koreans. *Journal of Neurolinguistics*, 17(1):79–91.

Vogt, P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3):429–457.

Vogt, P. (2005). On the Acquisition and Evolution of Compositional Languages: Sparse Input and the Productive Creativity of Children. *Adaptive Behavior*, 13(4):325–346.

Vogt, P. (2009). Modeling interactions between language evolution and demography. *Human Biology*, 81(2-3):237–58.

Vogt, P. and Divina, F. (2007). Social symbol grounding and language evolution. *Interaction Studies*, 8:31–52.

Xu, F. and Tenenbaum, J. B. (2000). Word learning as Bayesian inference. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 245–72, Austin, TX. Cognitive Science Society.

Xu, F. and Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3):288–97.

Xu, J., Griffiths, T. L., and Dowman, M. (2010). Replicating Color Term Universals through Human Iterated Learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

Yang, C. D. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10):451–6.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, volume 15, page 51, Cambridge (MA). The MIT Press.